Attorney Docket No.: 016336-001320US
Client Reference No.: 2478-3150-3442PT; 2478-3768-4270PT

PATENT APPLICATION

ANCESTRAL AND COT VIRAL SEQUENCES, PROTEINS AND IMMUNOGENIC COMPOSITIONS

Inventor(s): James I. Mullins, a citizen of The United States, residing at
3134 East Laurelhurst Drive N.E.
Seattle, WA 98105-5333

Allen G. Rodrigo, a citizen of New Zealand, residing at
8 Seaton Road, Murrays Bay
Auckland, New Zealand

Gerald H. Learn, a citizen of The United States, residing at
11316 N.E. 2nd Street
Kingston, WA 98346

Fusheng Li, a citizen of The United States, residing at
3818 N.E. 75th St., #3
Seattle, WA 98115

David C. Nickle, a citizen of The United States, residing at
5038 - 17th Ave. N.E.
Seattle, WA 98105

Mark A. Jensen, a citizen of The United States, residing at
15113 - 61st Ave. S.E.
Snohomish, WA 98296

Assignee: University of Washington
Office of Technology Licensing
4311 - 11th Avenue N.E., Suite 500, Campus Box 354990
Seattle, WA 98105-4608

Entity: Nonprofit Organization

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, Eighth Floor
San Francisco, California 94111-3834
Tel: 206-467-9600

ANCESTRAL AND COT VIRAL SEQUENCES, PROTEINS AND IMMUNOGENIC COMPOSITIONS

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application is a continuation-in-part and claims priority to U.S. Patent Application No. 10/204,204, filed February 16, 2001, which is a U.S. National Phase application of PCT/US01/05288, filed February 16, 2001, which claims the benefit of U.S. Provisional Patent Application No. 60/183,659, filed February 18, 2000. This application also claims the benefit of U.S. Provisional Application No. 60/447,586, filed February 14, 2003. All of these applications are incorporated by reference herein. All of these patent applications are incorporated herein by reference, in their entirety.

STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0002] This work was supported by a grant from the US Public Health Service through a grant to the University of Washington Center for AIDS Research (AI-27757) and PHS T32A107509 and T32CA0922925. The Federal Government may have certain rights in this invention.

BACKGROUND OF THE INVENTION

[0003] HIV-1 has proved to be an extremely difficult target for vaccine development. Immune correlates of protective immunity against HIV-1 infection remain uncertain. The virus persistently replicates in the infected individual, leading inexorably to disease despite the generation of vigorous humoral and cellular immune responses. HIV-1 rapidly mutates during infection, resulting in the generation of viruses that can escape immune recognition. Unlike other highly diverse viruses (*e.g.*, influenza), there does not appear to be a succession of variants where one prototypical strain is replaced by successive uniform strains. Rather, an evolutionary tree of viral sequences sampled from a large number of HIV-infected individuals form a star-burst pattern with most of the variants roughly equidistant from the center of the tree. HIV-1 viruses can also persist indefinitely as latent proviral DNA, capable of replicating in individuals at a later time.

[0004] Currently, several HIV-1 vaccine approaches are being developed, each with its own relative strengths and weaknesses. These approaches include the development of live attenuated vaccines, inactivated viruses with adjuvant peptides and subunit vaccines, live vector-based vaccines, and DNA vaccines. Envelope glycoproteins were considered as the prime antigen in the vaccine regimen due to their surface-exposure, until it became evident that they are not ideal immunogens. This is an expected consequence of the immunological selective forces that drive the evolution of these viruses: it appears that the same features of envelope glycoproteins that dictate poor immunogenicity in natural infections have hampered vaccine development. However, modification of the vaccine recipe may overcome these problems. For example, a recent report of successful neutralization (in mice) of primary isolates from infected individuals with a fusion-competent immunogen supports this idea.

[0005] Another approach could be to use natural isolates of HIV-1 in a vaccine recipe. Identification of early variants even from stored specimens near the start of the AIDS epidemic is very unlikely, however. Natural isolates are also unlikely to embody features (*e.g.*, epitopes) that are ideal for a vaccine candidate. Furthermore, any given natural virus isolate will have features that reflect adaptations due to specific interactions within that particular human host. These individual-specific features are not expected to be found in all or most strains of the virus, and thus vaccines based on individual isolates are unlikely to be effective against a broad range of circulating virus.

[0006] Another approach could be to include as many diverse HIV-1 isolates as possible in the vaccine recipe in an effort to elicit broad protection against HIV-1 challenge. First, one or more strains are chosen from among the many circulating strains of HIV. The advantage of this approach is that such a strain is known to be an infectious form of a viable virus. However, such a strain will be genetically quite dissimilar to other strains in circulation, and thus can fail to elicit broad protection. A related approach is to build a consensus sequence based on circulating strains, or on strains in the database. The consensus sequence is likely to be less distant in a genetic sense from circulating strains, but is not an estimate of any real virus, however, and thus may not provide broad protection.

[0007] Accordingly, there is a need in the art for new effective methods of identifying candidate sequences for vaccine development to prevent and treat HIV infection. The present invention fulfills this and other needs.

BRIEF SUMMARY OF THE INVENTION

[0008] The present invention provides methods for determining founder sequences from highly diverse virus populations. Also provided are determined founder sequences for highly diverse virus populations.

[0009] In one aspect, computational methods are provided for determining an ancestral viral sequence for highly diverse viruses, such as HIV-1, HIV-2 or Hepatitis C. These computational methods use samples of circulating viruses to determine an ancestral viral sequence by maximum likelihood phylogeny analysis. The ancestral viral sequence can be, for example, an HIV-1 ancestral viral gene sequence, an HIV-2 ancestral viral gene sequence, or a Hepatitis C ancestral viral gene sequence. In other embodiments, the ancestral viral gene sequence is of HIV-1 subtype A, B, C, D, E, F, G, H, J, AG, or AGI; HIV-1 Group M, N, or O; or HIV-2 subtype A or B. The ancestral viral gene sequence can be derived from widely dispersed HIV-1 variants, geographically-restricted HIV-1 variants, widely dispersed HIV-2 variants, or geographically-restricted HIV-2 variants. Typically, the ancestor gene is an *env* gene or a *gag* gene.

[0010] The ancestral viral gene sequence is more closely related, on average, to a gene sequence of any given circulating virus than to any other variant. In some embodiments, the ancestral viral gene sequence has at least 70% identity with the sequence set forth in SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, SEQ ID NO:6, but does not have 100% identity with any circulating viral variant.

[0011] In another aspect, ancestral sequences for the *env* gene of HIV-1 subtype B are provided. HIV-1 subtype B gives rise to most infections in the Western Hemisphere and in Europe. The determined ancestral viral sequence is on average more closely related to any given circulating virus than to any other variant. The *env* ancestral gene sequence encodes an open reading frame for gp160, the gene product of *env*. In additional specific embodiments, the ancestral viral sequence has a gene sequence set forth in Figures 9 to 17 (designated "mrca").

[0012] In another aspect, ancestral sequences for the *env* gene of HIV-1 subtype C are provided. Subtype C is the most prevalent subtype worldwide. The determined founder sequence is, on average, more closely related to any given circulating virus than to any other variant. This sequence encodes an open reading frame for gp160, the gene product of *env*.

In additional specific embodiments, the ancestral viral sequence has a gene sequence set forth in Figures 27 to 35 (designated "mrca").

[0013] Isolated HIV ancestor proteins or fragments thereof are also provided. The isolated ancestor protein can be, for example, the contiguous sequence of HIV-1, subtype B, *env* ancestor protein (SEQ ID NO:2) or HIV-1, subtype C, *env* ancestor protein (SEQ ID NO:4). The ancestor protein can also be of HIV-1 subtype A, B, C, D, E, F, G, H, J, AG, or AGI; HIV-1 Group M, N, or O; or HIV-2 subtype A or B. In additional specific embodiments, the ancestor protein can have a sequence set forth in Figures 18-26 or 36 to 44 (designated "mrca").

[0014] In another aspect, computational methods are provided for determining a COT viral sequence for highly diverse viruses, such as HIV-1, HIV-2 or Hepatitis C. These computational methods use samples of circulating viruses to determine a COT viral sequence by Least Squares or Minimum of Means methodologies. The COT viral sequence can be, for example, an HIV-1 COT viral gene sequence, an HIV-2 COT viral gene sequence, or a Hepatitis C COT viral gene sequence. In other embodiments, the COT viral gene sequence is of HIV-1 subtype A, B, C, D, E, F, G, H, J, AG, or AGI; HIV-1 Group M, N, or O; or HIV-2 subtype A or B. The COT viral gene sequence can be derived from widely dispersed HIV-1 variants, geographically-restricted HIV-1 variants, widely dispersed HIV-2 variants, or geographically-restricted HIV-2 variants. Typically, the COT viral gene is an *env* gene or a *gag* gene.

[0015] The COT viral gene sequence is more closely related, on average, to a gene sequence of any given circulating virus than to any other variant. In certain embodiments, the COT viral gene sequence has at least 70% identity with the LScot and MMcot sequences set forth in Figures 9 to 17 or 27 to 35, but does not have 100% identity with any circulating viral variant.

[0016] In another aspect, COT sequences for genes of HIV-1 subtype B are provided. HIV-1 subtype B gives rise to most infections in the Western Hemisphere and in Europe. The determined COT viral sequences are, on average, more closely related to any given circulating virus than to any other variant. In specific embodiments, the COT viral gene sequence is an LScot or MMcot sequence set forth in Figures 9 to 17, but does not have 100% identity with any circulating viral variant.

[0017] In another aspect, COT sequences for genes of HIV-1 subtype C are provided. Subtype C is the most prevalent subtype worldwide. The determined COT sequence is, on average, more closely related to any given circulating virus than to any other variant. In specific embodiments, the COT viral gene sequence is an LScot or MMcot sequence set forth in Figures 27 to 35, but does not have 100% identity with any circulating viral variant.

[0018] Isolated HIV COT proteins and or fragments thereof are also provided. The isolated ancestor protein can be, for example, an LScot or MMcot amino acid sequence set forth in Figures 27 to 25 and 36 to 44.

[0019] Also provided are computational methods for determining other ancestral or COT viral sequences. The computational methods can be extended, for example, to determine an ancestral or COT viral sequence for other HIV subtypes, such as, for example, HIV-1 subtype E, which is widely spread in developing countries. The computational methods can also be extended to determine an ancestral or COT viral sequence for all known and newly emerging highly diverse virus, such as, for example, HIV-1 strains, subtypes and groups. For example, ancestral or COT viral sequences can be determined for HIV-1-B in Thailand or Brazil, HIV-1-C in China, India, South Africa or Brazil, and the like. In other embodiments, the ancestral or COT viral sequence is determined for the HIV-1 *nef* gene or polypeptide, *pol* gene or polypeptide or other auxiliary genes or polypeptide (*see infra*).

[0020] The present invention also provides an expression construct including a transcriptional promoter; a nucleic acid encoding an ancestor protein; and a transcriptional terminator. The nucleic acid can encode, for example, an HIV-1 ancestor or COT protein (*e.g.*, SEQ ID NO:2 or SEQ ID NO:4). The nucleic acid can be, for example, an HIV-1 subtype B or C *env* gene sequence (*e.g.*, SEQ ID NO:1, SEQ ID NO:3, SEQ ID NO:5, or SEQ ID NO:6). In one embodiment, the nucleic acid sequence is optimized for expression in a host cell.

[0021] The promoter can be a heterologous promoter, such as the cytomegalovirus promoter. The expression construct can be expressed in prokaryotic or eukaryotic cells. Suitable cells include, for example, mammalian cells, human cells, *Escherichia coli* cells, and *Saccharomyces cerevisiae* cells. In one embodiment, the expression construct has the nucleic acid sequence operably linked to a Semliki Forest Virus replicon, wherein the resulting recombinant replicon is operably linked to a cytomegalovirus promoter.

[0022] In another aspect, compositions are provided for inducing an immune response in a mammal, the compositions include a viral ancestor protein or an immunogenic fragment of an ancestor or COT protein. The ancestor or COT protein can be derived from HIV-1 subtype B or C *env* ancestor or COT protein, or from other HIV-1, HIV-2 or Hepatitis C ancestor or COT proteins. In other aspects, the composition can be used as a vaccine, such as an AIDS vaccine to protect against infection by the highly diverse human immunodeficiency virus, type 1 (HIV-1), or for protection against HIV-2 or Hepatitis C infections. The ancestral or COT viral sequence can be an HIV-1 group determined founder (*e.g.*, for Group M), for an HIV-1 subtype (*e.g.*, B, C or E), for a widely spread variant, for a geographically-restricted variant, or for a newly emerging variant.

[0023] In another aspect, isolated antibodies are provided that bind specifically to a viral ancestor or COT protein and that bind specifically to a plurality of circulating descendant viral ancestor or COT proteins. The ancestor or COT protein can be from, for example, HIV-1, HIV-2, or Hepatitis C. The antibody can be a monoclonal antibody or antigen binding fragment thereof. In one embodiment, the antibody is a humanized monoclonal antibody. Other suitable antibodies or antigen binding fragments thereof can be a single chain antibody, a single heavy chain antibody, an antigen binding $F(ab')_2$ fragment, an antigen binding Fab' fragment, an antigen binding Fab fragment, or an antigen binding Fv fragment.

[0024] In addition to determining ancestral and COT viral sequences, the present invention also provides methods for preparing and testing immunogenic compositions based on an ancestral or COT viral sequence. In specific embodiments, immunogenic compositions (based on an ancestral or COT viral sequence) are prepared and administered to a mammal, employing an appropriate model, such as, for example, a mouse model or simian-human immunodeficiency virus (SHIV) macaque model. Immunogenic compositions can be prepared using an isolated ancestral or COT viral gene sequence, or polypeptide sequence, or a portion thereof.

[0025] In yet another aspect, diagnostic methods are provided to detect HIV and/or AIDS in a subject, using the nucleic acids, peptides or antibodies based on an ancestral or COT viral sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

[0026] Figure 1 shows a phylogenetic classification of HIV-1. The circled nodes approximate the ancestral state of the HIV-1 main group (Group M) and the main group clades A-G, J, AGI and AG.

[0027] Figure 2 shows the phylogenetic relationship of HIV-1 subtype B and the placement of the determined subtype B ancestral node on that tree. The phylogenetic relationship of HIV-1 subtype D is shown as an outgroup.

[0028] Figure 3 shows an ancestral viral sequence reconstruction of the most recent common ancestor using maximum likelihood reconstruction for an SIV inoculum up to three years after infection into macaques. The consensus sequence and the most recent common ancestor sequence were found to differ 1.5% in nucleotide sequence.

[0029] Figure 4 provides an example of the development of a digital vaccine using an ancestral viral sequence.

[0030] Figure 5 shows a comparison of a "most parsimonious reconstruction" methodology and a "maximum likelihood reconstruction methodology."

[0031] Figure 6 shows another comparison of the "most parsimonious reconstruction" methodology and the "maximum likelihood reconstruction methodology."

[0032] Figure 7 illustrates a map of the pJW4304 SV40/EBV vector.

[0033] Figure 8 shows the phylogenetic relationship of HIV-1 subtype C and the placement of the determined subtype C ancestral node on that tree.

[0034] Figure 9 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *gag* gene.

[0035] Figure 10 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *env* (encoding gp160).

[0036] Figure 11 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *nef* gene.

[0037] Figure 12 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *pol* gene.

[0038] Figure 13 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *rev* gene.

[0039] Figure 14 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *tat* gene.

[0040] Figure 15 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *vif* gene.

[0041] Figure 16 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *vpr* gene.

[0042] Figure 17 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B *vpu* gene.

[0043] Figure 18 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B gag protein.

[0044] Figure 19 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B gp160 protein.

[0045] Figure 20 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B nef protein.

[0046] Figure 21 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B pol protein.

[0047] Figure 22 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B rev protein.

[0048] Figure 23 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B tat protein.

[0049] Figure 24 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B vif protein.

[0050] Figure 25 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B vpr protein.

[0051] Figure 26 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade B vpu protein.

[0052] Figure 27 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *gag* gene.

[0053] Figure 28 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *env* (encoding gp160).

[0054] Figure 29 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *nef* gene.

[0055] Figure 30 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *pol* gene.

[0056] Figure 31 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *rev* gene.

[0057] Figure 32 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *tat* gene.

[0058] Figure 33 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *vif* gene.

[0059] Figure 34 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *vpr* gene.

[0060] Figure 35 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C *vpu* gene.

[0061] Figure 36 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C gag protein.

[0062] Figure 37 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C gp160 protein.

[0063] Figure 38 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C nef protein.

[0064] Figure 39 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C pol protein.

[0065] Figure 40 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C rev protein.

[0066] Figure 41 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C tat protein.

[0067] Figure 42 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C vif protein.

[0068] Figure 43 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C vpr protein.

[0069] Figure 44 shows a comparison of the Most Recent Common Ancestor ("MRCA"), COT Least Squares ("LScot") and COT Minimum of Means ("MMcot") reconstructions for the Clade C vpu protein.

[0070] Figure 45. Phylogenetic relationships of different phylogenetic structures and between HIV-1 group M, Subtype B gp160 sequences. A. thirty-eight Subtype B sequences and three Subtype D (outgroup) sequences used to root the Subtype B sequences (*see* Table 11). The Subtype B sequences were from nine countries, representing a broad sample of Subtype B diversity: Australia, 8; China, 1; France, 6; Gabon, 1; Germany, 2; Great Britain, 2; The Netherlands, 2; Spain, 1; USA, 15). B. Idealized phylogenetic trees with caterpillar (left) and star (right) shapes.

[0071] Figure 46. Deduced ancestor protein sequences SIVBK28 ancestor (Env segment) and AN1-EnvB.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

[0072] The present invention provides methods for determining founder sequences from highly diverse virus populations. Also provided are determined founder sequences for highly diverse virus populations.

[0073] Prior to setting forth the invention in more detail, it may be helpful to a further understanding thereof to set forth definitions of certain terms as used hereinafter.

[0074] *Definitions*

[0075] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention pertains. Although any methods and materials similar to those described herein can be used in the practice or testing of the present invention, only exemplary methods and

materials are described. For purposes of the present invention, the following terms are defined below.

[0076] An "ancestral sequence" refers to a determined founder sequence, determined through application of maximum likelihood phylogenetic analysis. An ancestral sequence is typically one that is more closely related, on average, to any given variant than to any other variant. An "ancestral viral sequence" refers to a determined founder sequence, typically one that is more closely related, on average, to any given circulating virus than to any other variant. An "ancestral viral sequence" is determined through application of maximum likelihood phylogenetic analysis using the nucleic acid and/or amino acid sequences of circulating viruses. An "ancestor virus" is a virus comprising the "ancestral viral sequence." An "ancestor protein" is a protein, polypeptide or peptide having an amino acid ancestral viral sequence.

[0077] A "COT sequence" refers to a determined founder sequence, determined through application of a COT Least Squares Method or a COT Minimum of Means Method. A "COT sequence" is a position at a node or on a branch of a phylogenetic tree having completely specified branch lengths. A "COT viral sequence" refers to a founder nucleic acid sequence determined by COT Least Squares Method or a COT Minimum of Means Method, using the nucleic acid and/or amino acid sequences of circulating viruses. An "COT virus" is a virus comprising the "COT viral sequence." A "COT viral protein" or "COT protein" is a protein, polypeptide or peptide having an amino acid COT viral sequence.

[0078] The term "circulating virus" refers to virus found in an infected individual.

[0079] The term "variant" refers to a virus, gene or gene product that differs in sequence from other viruses, genes or gene products by one or more nucleotide or amino acids.

[0080] The terms "immunological" or "immune response" refer to the development of a beneficial humoral (*i.e.*, antibody mediated) and/or a cellular (*i.e.*, mediated by antigen-specific T-cells or their secretion products) response directed against an HIV peptide in a recipient subject. Such a response can be, in particular, an active response induced by the administration of an immunogen. A cellular immune response is elicited by the presentation of epitopes in association with Class I or Class II MHC molecules to activate antigen-specific $CD4^+$ T helper cells (*i.e.*, Helper T lymphocytes) and/or $CD8^+$ cytotoxic T cells. The presence of a cell-mediated immunological response can be determined by, for example, proliferation assays of $CD4^+$ T cells (*i.e.*, measuring the HTL (Helper T lymphocyte)

response) or by CTL (cytotoxic T lymphocyte) assays (*see, e.g.*, Burke *et al.*, *J. Inf. Dis.* 170:1110-19 (1994); Tigges *et al.*, *J. Immunol.* 156:3901-10 (1996)). The relative contributions of humoral and cellular responses to the protective or therapeutic effect of an immunogen can be distinguished by separately isolating IgG and T-cells from an immunized syngeneic animal and measuring protective or therapeutic effects in a second subject. For example, the effector cells can be deleted and the resulting response analyzed (*see, e.g.*, Schmitz *et al.*, *Science* 283:857-60 (1999); Jin *et al.*, *J Exp. Med.* 189:991-98 (1999)).

[0081] "Antibody" refers to a polypeptide substantially encoded by an immunoglobulin gene or immunoglobulin genes, or fragments thereof, that specifically bind and recognize an analyte (antigen). The recognized immunoglobulin genes include the kappa, lambda, alpha, gamma, delta, epsilon and mu constant region genes, as well as the myriad immunoglobulin variable region genes. Light chains are classified as either kappa or lambda. Heavy chains are classified as gamma, mu, alpha, delta, or epsilon, which in turn define the immunoglobulin classes, IgG, IgM, IgA, IgD and IgE, respectively.

[0082] An exemplary immunoglobulin (antibody) structural unit comprises a tetramer. Each tetramer is composed of two identical pairs of polypeptide chains, each pair having one "light" (about 25 kD) and one "heavy" chain (about 50-70 kD). The N-terminus of each chain has a variable region of about 100 to 110 or more amino acids primarily responsible for antigen recognition. The terms variable light chain (VL) and variable heavy chain (VH) refer to these light and heavy chains, respectively.

[0083] Antibodies exist, for example, as intact immunoglobulins or as a number of well characterized antigen-binding fragments produced by digestion with various peptidases. For example, pepsin digests an antibody below the disulfide linkages in the hinge region to produce an $F(ab')_2$ fragment, a dimer of Fab which itself is a light chain joined to VH-CH1 by a disulfide bond. The $F(ab')_2$ fragment can be reduced under mild conditions to break the disulfide linkage in the hinge region, thereby converting the $F(ab')_2$ dimer into an Fab' monomer. The Fab' monomer is essentially an Fab with part of the hinge region (*see*, *Fundamental Immunology*, Third Edition, W.E. Paul (ed.), Raven Press, N.Y. (1993)). While various antibody fragments are defined in terms of the digestion of an intact antibody, one of skill will appreciate that such fragments can be synthesized *de novo* either chemically or by utilizing recombinant DNA methodology. Thus, the term antibody, as used herein, also includes antibody fragments, such as a single chain antibody, an antigen binding $F(ab')_2$

fragment, an antigen binding Fab' fragment, an antigen binding Fab fragment, an antigen binding Fv fragment, a single heavy chain or a chimeric antibody. Such antibodies can be produced by the modification of whole antibodies or synthesized *de novo* using recombinant DNA methodologies.

[0084] The term "biological sample" refers to any tissue or liquid sample having genomic or viral DNA or other nucleic acids (*e.g.*, mRNA, viral RNA, or the like) or proteins. "Biological sample" further includes fluids, such as serum and plasma, that contain cell-free virus, and also includes both normal healthy cells and cells suspected of HIV infection.

[0085] The term "nucleic acid" refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single or double stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides that have similar binding properties as the reference nucleic acid. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions) and complementary sequences as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions can be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (*see, e.g.*, Batzer *et al.*, *Nucleic Acid Res.* 19:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* 260:2605-08 (1985); Rossolini *et al.*, *Mol. Cell. Probes* 8:91-98 (1994)). Nucleic acids also include fragments of at least 10 contiguous nucleotides (*e.g.*, a hybridizable portion); in other embodiments, the nucleic acids comprise at least 25 nucleotides, 50 nucleotides, 100 nucleotides, 150 nucleotides, 200 nucleotides, or even up to 250 nucleotides or more. The term "nucleic acid" is used interchangeably with gene, cDNA, and mRNA encoded by a gene.

[0086] As used herein a "nucleic acid probe" is defined as a nucleic acid capable of binding to a target nucleic acid (*e.g.*, an HIV-1 nucleic acid) of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, such as by hydrogen bond formation. As used herein, a probe may include natural (*e.g.*, A, G, C, or T) or modified bases (*e.g.*, 7-deazaguanosine, inosine, *or the like*). In addition, the bases in a probe can be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, for example, probes can be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages. It will be understood by one of skill in the art that probes can bind target sequences lacking

complete complementarity with the probe sequence, at levels that depend upon the stringency of the hybridization conditions.

[0087] Nucleic acid probes can be DNA or RNA fragments. DNA fragments can be prepared, for example, by digesting plasmid DNA, by use of PCR, or by chemical synthesis, such as by the phosphoramidite method described by Beaucage and Carruthers (*Tetrahedron Lett.* 22:1859-62 (1981)), or by the triester method according to Matteucci *et al.* (*J. Am. Chem. Soc.* 103:3185 (1981)). A double stranded fragment can then be obtained, if desired, by annealing the chemically synthesized single strands together under appropriate conditions, or by synthesizing the complementary strand using DNA polymerase with an appropriate primer sequence. Where a specific sequence for a nucleic acid probe is given, it is understood that the complementary strand is also identified and included. The complementary strand will work equally well in situations where the target is a double stranded nucleic acid.

[0088] A "labeled nucleic acid probe" is a nucleic acid probe that is bound, either covalently, through a linker, or through ionic, van der Waals or hydrogen bonds, to a label such that the presence of the probe can be detected by detecting the presence of the label bound to the probe.

[0089] The term "operably linked" refers to functional linkage between a nucleic acid expression control sequence (such as a promoter, signal sequence, or any of an array of transcription factor binding sites) and a second nucleic acid sequence, wherein the expression control sequence affects transcription and/or translation of the nucleic acid corresponding to the second sequence.

[0090] "Amplification primers" are nucleic acids, typically oligonucleotides, comprising either natural or analog nucleotides that can serve as the basis for the amplification of a selected nucleic acid sequence. They include, for example, both polymerase chain reaction primers and ligase chain reaction oligonucleotides.

[0091] The terms "polypeptide," "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymers.

[0092] The terms "amino acid" or "amino acid residue", as used herein, refer to naturally occurring L-amino acids or to D-amino acids as described further below. The commonly used one- and three-letter abbreviations for amino acids are used herein (*see*, *e.g.*, Alberts *et al.*, *Molecular Biology of the Cell*, Garland Publishing, Inc., New York (3d ed. 1994); Creighton, *Proteins*, W.H. Freeman and Company (1984)).

[0093] A "conservative substitution," when describing a protein, refers to a change in the amino acid composition of the protein that is less likely to substantially alter the protein's activity. Thus, "conservatively modified variations" of a particular amino acid sequence refers to amino acid substitutions of those amino acids that are less likely to be critical for protein activity or substitution of amino acids with other amino acids having similar properties (*e.g.*, acidic, basic, positively or negatively charged, polar or non-polar, or the like) such that the substitutions of even critical amino acids do not substantially alter activity. Conservative substitution tables providing amino acids that are often functionally similar are well known in the art (*see*, *e.g.*, Creighton, *Proteins*, W.H. Freeman and Company (1984)). In addition, individual substitutions, deletions or additions which alter, add or delete a single amino acid or a small percentage of amino acids in an encoded sequence are also "conservatively modified variations."

[0094] The terms "identical" or "percent identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same (*i.e.*, 60% identity, optionally 65%, 70%, 75%, 80%, 85%, 90%, or 95% identity over a specified region), when compared and aligned for maximum correspondence over a comparison window, or designated region, as measured using one of the following sequence comparison algorithms or by manual alignment and visual inspection. Such sequences are then said to be "substantially identical." This definition also refers to the complement of a test sequence. Optionally, the identity exists over a region that is at least about 30 amino acids or nucleotides in length, typically over a region that is 50, 75 or 150 amino acids or nucleotides. In certain embodiments, the sequences are substantially identical over the entire length of the coding regions.

[0095] The terms "similarity," or "percent similarity," in the context of two or more polypeptide sequences, refer to two or more sequences or subsequences that have a specified percentage of amino acid residues that are either the same or similar as defined in the

conservative amino acid substitutions defined above (*i.e.*, at least 60%, optionally 65%, 70%, 75%, 80%, 85%, 90%, or 95% similar over a specified region), when compared and aligned for maximum correspondence over a comparison window, or designated region as measured using one of the following sequence comparison algorithms or by manual alignment and visual inspection. Such sequences are then said to be "substantially similar." Optionally, this identity exists over a region that is at least about 25 amino acids in length, or more typically over a region that is at least about 50, 75 or 100 amino acids in length.

[0096] For sequence comparison, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are typically input into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated program parameters.

[0097] Optimal alignment of sequences for comparison can be conducted, for example, by the local homology algorithm of Smith and Waterman (*Adv. Appl. Math.* 2:482 (1981)), by the homology alignment algorithm of Needleman and Wunsch (*J. Mol. Biol.* 48:443 (1970)), by the search for identity method of Pearson and Lipman (*Proc. Natl. Acad. Sci. USA* 85:2444 (1988)), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by visual inspection (*see*, *generally* Ausubel *et al.*, *Current Protocols in Molecular Biology*, John Wiley and Sons, New York (1996)).

[0098] One example of a useful algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments to show relationship and percent sequence identity. It also plots a tree or dendogram showing the clustering relationships used to create the alignment. PILEUP uses a simplification of the progressive alignment method of Feng and Doolittle (*J. Mol. Evol.* 35:351-60 (1987)). The method used is similar to the CLUSTAL method described by Higgins and Sharp (*Gene* 73:237-44 (1988); *CABIOS* 5:151-53 (1989)). The program can align up to 300 sequences, each of a maximum length of 5,000 nucleotides or amino acids. The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster is then aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences are aligned by a simple

extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments. The program is run by designating specific sequences and their amino acid or nucleotide coordinates for regions of sequence comparison and by designating the program parameters. For example, a reference sequence can be compared to other test sequences to determine the percent sequence identity relationship using the following parameters: default gap weight (3.00), default gap length weight (0.10), and weighted end gaps.

[0099] Another example of an algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al.* (*J. Mol. Biol.* 215:403-10 (1990)). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always> 0) and N (penalty score for mismatching residues; always <0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff and Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)).

[0100] In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see*, *e.g.*, Karlin and

Altschul, *Proc. Natl. Acad. Sci. USA* 90:5873-87 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is typically between about 0.35 and about 0.1. Another indication that two nucleic acids are substantially identical is that the two molecules hybridize to each other under stringent conditions. The phrase "hybridizing specifically to" refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. "Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

[0101] "Stringent hybridization conditions" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments, such as Southern and northern hybridizations, are sequence-dependent, and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes*, part I, chapter 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays," Elsevier, N.Y. (1993). Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Typically, under "stringent conditions," a probe will hybridize to its target subsequence, but to no other sequences.

[0102] The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe. An example of stringent hybridization conditions for hybridization of complementary nucleic acids which have more than 100 complementary residues on a filter in a Southern or northern blot is 50% formamide in 4-6x SSC or SSPE at 42°C, or 65-68° C in aqueous solution containing 4-6x SSC or SSPE. An example of highly stringent wash conditions is 0.15 M NaCl at 72°C for about 15 minutes. An example of stringent wash conditions is a 0.2X SSC wash at 65°C for 15 minutes. (*See*

generally Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual*, 2nd ed., Cold Spring Harbor Publish., Cold Spring Harbor, NY (1989)). Often, a high stringency wash is preceded by a low stringency wash to remove background probe signal. An example of medium stringency wash for a duplex of, for example, more than 100 nucleotides, is 1X SSC at 45°C for 15 minutes. An example of low stringency wash for a duplex of, for example, more than 100 nucleotides, is 4-6X SSC at 40°C for 15 minutes. For short probes (*e.g.*, about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.0 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30°C. Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2X (or higher) than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization. Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides which they encode are substantially identical. This occurs, for example, when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

[0103] A further indication that two nucleic acids or polypeptides are substantially identical is that the polypeptide encoded by the first nucleic acid is immunologically cross reactive with, or specifically binds to, antibodies raised against the polypeptide encoded by the second nucleic acid. Thus, a polypeptide is typically substantially identical to a second polypeptide, for example, where the two peptides differ only by conservative substitutions.

[0104] The phrase "specifically (or selectively) binds to an antibody" or "specifically (or selectively) immunoreactive with", when referring to a protein or peptide, refers to a binding reaction which is determinative of the presence of the protein in the presence of a heterogeneous population of proteins and other biologics. Thus, under designated immunoassay conditions, the specified antibodies bind to a particular protein and do not bind in a significant amount to other proteins present in the sample. Specific binding to a protein under such conditions may require an antibody that is selected for its specificity for the particular protein. For example, antibodies raised to the protein with the amino acid sequence encoded by any of the nucleic acids of the invention can be selected to obtain antibodies specifically immunoreactive with that protein and not with other proteins except for polymorphic variants. A variety of immunoassay formats can be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA

immunoassays, Western blots, or immunohistochemistry are routinely used to select monoclonal antibodies specifically immunoreactive with a protein (*see*, *e.g.*, Harlow and Lane, *Antibodies, A Laboratory Manual*, Cold Spring Harbor Publications, N.Y. (1988), for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity). Typically, a specific or selective reaction will be at least twice background signal or noise and more typically more than 10 to 100 times background.

[0105] The term "immunogenic composition" refers to a composition that elicits an immune response which produces antibodies or cell-mediated immune responses against a specific immunogen. Immunogenic compositions can be prepared as injectables, as liquid solutions, suspensions, emulsions, and the like.

[0106] The term "vaccine" refers to an immunogenic composition for *in vivo* administration to a host, which may be a primate, particularly a human host, to confer protection against disease, particularly a viral disease.

[0107] The term "isolated" refers to a virus, nucleic acid or polypeptide that has been removed from its natural cellular environment. An isolated virus, nucleic acid or polypeptide is typically at least partially purified from cellular nucleic acids, polypeptides and other constituents.

[0108] In the context of the present invention, a "Coalescent Event" refers to the joining of two lineages on a genealogy at the point of their most recent common ancestor.

[0109] A "Coalescent Interval" describes the time between coalescent events. The expected time for each coalescent interval is exponentially distributed with mean $E\,[t_{nyn\text{-}1}] = 2N / n\,(n - 1)$ generations for $n << N$.

[0110] *Determination of Ancestral Sequences*

[0111] In one aspect, computational methods are provided for determining ancestral sequences. Such methods can be used, for example, to determine ancestral sequences for viruses. These computational methods are typically used to determine an ancestral sequence of a virus that exists as a highly diverse viral population. For example, some highly diverse viruses (including HIV-1, HIV-2, Hepatitis C, and the like) do not appear to evolve through a succession of variants, where one prototypical strain is replaced by successive uniform strains. Instead, an evolutionary tree of viral sequences can form a "star-burst pattern," with

most of the variants approximately equidistant from the center of the star-burst. This star-burst pattern indicates that multiple, diverse circulating strains evolve from a common ancestor. The computational methods can be used to determine ancestral sequences for such highly diverse viruses, such as, for example, HIV-1, HIV-2, Hepatitis C, and other viruses.

[0112] Methods for determining ancestral sequences are typically based on the nucleic acid sequences of circulating viruses. As a viral nucleic acid sequence is replicated, it acquires base changes due to errors in the replication process. For example, as some nucleic acid sequences are replicated, thymine (T) might bind to a guanine (G) rather than its normal complement, cytosine (C). Most of these base changes (or mutations) are not reproduced in subsequent replication events, but a certain proportion of mutations are passed down to the descendant sequences. With more replication cycles, nucleic acid sequences acquire more mutations. If a nucleic acid sequence bearing one or more mutations gives rise to two separate lineages, then the resulting two lineages will share the same parental nucleic acid sequence, and have the same parental mutation(s). If the "histories" of these lineages are traced backwards, they will have a common branch point, at which the two lineages arose from a common ancestor. Similarly, if the histories of presently circulating viral nucleic acid sequences are traced backwards, the branching points in these histories also correspond to points, designated as nodes, at which a single ancestor gave rise to the descendant lineages.

[0113] The present computational methods are based on the principle of maximum likelihood and use samples of nucleic acid sequences of circulating viruses. The sequences of the viruses in the samples typically share a common feature, such as being from the same viral strain, subtype or group. A phylogeny is constructed by using a model of evolution that specifies the probabilities of nucleotide substitutions in the replicating viral nucleic acids. At positions in the sequences where the nucleotides differ (*i.e.*, at the site of a mutation), the methodology assigns one of the nucleotides to the node (*i.e.*, the branch point of the lineages) such that the probability of obtaining the observed viral sequences is maximized. The assignment of nucleotides to the nodes is based on the predicted phylogeny or phylogenies. For each data set, several sequences from a different viral strain, subtype or group are used as an outgroup to root the sequences of interest. A model of sequence substitutions and then a maximum likelihood phylogeny are determined for each data set (*e.g.*, subtype and outgroup). The maximum likelihood phylogeny the one that has the highest probability of giving the observed nucleic acid sequences in the samples. The sequence at the base node of the maximum likelihood phylogeny is referred to as the ancestral sequence (or most recent

common ancestor). (*See, e.g.*, Figures 1 and 2). This ancestral sequence is thus approximately equidistant from the different sequences within the samples.

[0114] Maximum likelihood phylogeny uses samples of the sequences of circulating virus. The sequences of circulating viruses can be determined, for example, by extracting nucleic acids from blood, tissues or other biological samples of virally infected persons and sequencing the viral nucleic acids. (*See, e.g.*, Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual*, 2nd ed., Cold Spring Harbor Publish., Cold Spring Harbor, N.Y. (1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual*, W.H. Freeman, N.Y. (1990); Ausubel *et al.*, *supra.*) In one embodiment, extracted viral nucleic acids can be amplified by polymerase chain reaction, and then DNA sequenced. Samples of circulating virus can be obtained from stored biological samples and/or prospectively from samples of circulating virus (*e.g.*, sampling HIV-1 subtype C in India versus Ethiopia). Viral sequences can also be identified from databases (*e.g.*, GenBank and Los Alamos sequence databases).

[0115] Once samples of circulating viruses are collected (typically about 20 to about 50 samples), the nucleic acid sequences for one or more genes are analyzed using the computational methods according to the present invention. In one method, for any given site in the sequence, the nucleotides at all nodes on a tree are assigned. The configuration of the nucleotides for all nodes that maximizes the probability of obtaining the observed sequences of circulating viruses is determined. With this method, the joint likelihood of the states across all nodes is maximized.

[0116] A second method is to choose, for a given nucleotide site and a given node on the tree, the nucleotide that maximizes the probability of obtaining the observed sequences of circulating viruses, allowing for all possible assignments of nucleotides at the other nodes on the tree. This second method maximizes the marginal likelihood of a particular assignment. For these methods, the reconstruction of the ancestral sequence (*i.e.*, ancestral state) need not result in only a single determined sequence, however. It is possible to choose a number of ancestral sequences, ranked in order of their likelihood.

[0117] With HIV populations, a second layer of modeling can be added to the maximum likelihood phylogenetic analysis, in particular the layer is added to the model of evolution that is employed in the analysis. This second layer is based on coalescent likelihood analysis. The coalescent is a mathematical description of a genealogy of sequences, taking account of the processes that act on the population. If these processes are known with some certainty,

the use of the coalescent can be used to assign prior probabilities to each type of tree. Taken together with the likelihood of the tree, the posterior probability can be determined that a determined phylogenetic tree is correct given the data. Once a tree is chosen, the ancestral states are determined, as described above. Thus, coalescent likelihood analysis can also be applied to determine the sequence of an ancestral viral sequence (*e.g.*, a founder, or Most Recent Common Ancestor (MRCA), sequence).

[0118] In a typical embodiment, maximum likelihood phylogeny analysis is applied to determine an ancestor sequence (*e.g.*, an ancestral viral sequence). Typically, between 20 and 50 nucleic acid sequence samples are used that have a common feature, such as a viral strain, subtype or group (*e.g.*, samples encompassing a worldwide diversity of the same subtype). Additional sequences from other viruses (*e.g.*, another strain, subtype, or group) are obtained and used as an outgroup to root the viral sequences being analyzed. The samples of viral sequences are determined from presently circulating viruses, identified from the database (*e.g.*, GenBank and Los Alamos sequence databases), or from similar sources of sequence information. The sequences are aligned using CLUSTALW (Thompson *et al.*, *Nucleic Acids Res.* 22:4673-80 (1994), the disclosure of which is incorporated by reference herein) and these alignments are refined using GDE (Smith *et al.*, *CABIOS* 10:671-75 (1994) the disclosure of which is incorporated by reference herein). The amino acid sequences are also translated from the nucleic acid sequences. Gaps are manipulated so that they are inserted between codons. This alignment (alignment I) is modified for phylogenetic analysis so that regions that can not be unambiguously aligned are removed (Learn *et al.*, *J. Virol.* 70:5720-30 (1996), the disclosure of which is incorporated by reference herein) resulting in alignment II.

[0119] An appropriate evolutionary model for phylogeny and ancestral state reconstructions for these sequences (alignment II) is selected using the Akaike Information Criterion (AIC) (Akaike, *IEEE Trans. Autom. Contr.* 19:716-23 (1974); which is incorporated by reference herein) as implemented in Modeltest 3.0 (Posada and Crandall, *Bioinformatics* 14:817-8 (1998), which is incorporated by reference herein). For example, for the analysis for the subtype C ancestral sequence the optimal model is equal rates for both classes of transitions and different rates for all four classes of transversions, with invariable sites and a X distribution of site-to-site rate variability of variable sites (referred to as a TVM+I+G model). The parameters of the model in this case can be, for example, equilibrium nucleotide frequencies: $f_A = 0.3576, f_C = 0.1829, f_G = 0.2314, f_T = 0.2290$;

proportion of invariable sites = 0.2447; shape parameter (α) of the X distribution = 0.7623; rate matrix (R) matrix values: $R_{A\rightarrow C} = 1.7502$, $R_{A\rightarrow G} = R_{C\rightarrow T} = 4.1332$, $R_{A\rightarrow T} = 0.6825$, $R_{C\rightarrow G} = 0.6549$, $R_{G\rightarrow T} = 1$.

[0120] Evolutionary trees for the sequences (alignment II) are inferred using maximum likelihood estimation (MLE) methods as implemented in PAUP* version 4.0b (Swofford, PAUP 4.0: Phylogenetic Analysis Using Parsimony (And Other Methods); Sinauer Associates, Inc. (2000) the disclosure of which is incorporated by reference herein). For example, for HIV-1 subtype C sequences, ten different subtree-pruning-regrafting (SPR) heuristic searches can be performed, each using a different random addition order. The ancestral viral nucleotide sequence is determined to be the sequence at the basal node using the phylogeny, the sequences from the databases (alignment II), and the TVM+I+G model above using marginal likelihood estimation (*see infra*).

[0121] In some cases, the determined sequence may not include ancestral sequence for portions of variable regions (*e.g.*, variable regions V1, V2, V4 and V5 for HIV-1-C), and or some short regions may not be unambiguously aligned. The following procedure can optionally be used to predict amino acid sequences for the complete sequence, including the highly variable regions (such as those deleted from alignment I). The determined ancestral sequence is visually aligned to alignment I and translated using GDE (Smith *et al.*, *supra*). Since the highly variable regions can be deleted as complete codons, the translational reading frame can be preserved and codons can be maintained. The ancestral amino acid sequence for the regions deleted from alignment II can be predicted visually and refined using a parsimony-based sequence reconstruction for these sites using the computer program MacClade, version 3.08a (Maddison and Maddison. MacClade — Analysis of Phylogeny and Character Evolution — Version 3. Sinauer Associates, Inc. (1992)).

[0122] The ancestral amino acid sequence is optionally optimized for expression in a particular cell type. Amino acid sequences can converted to a DNA sequence optimized for expression in certain cell types (*e.g.*, human cells) using, for example, the BACKTRANSLATE program of the Wisconsin Sequence Analysis Package (GCG), version 10 and a human gene codon table from the Codon Usage Database (www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=Homo+sapiens+[gbpri]), both incorporated by reference herein.

[0123] The optimized sequences encode the same amino acid sequence for the gene of interest (*e.g.*, the *env* gene) as the non-optimized ancestral sequence. A synthetic virus having the optimized sequence may not be fully functional due to the disruption of auxiliary genes in different reading frames the presence of RNA secondary structural feature (*e.g.*, the Rev responsive element (RRE) of HIV-1), and the like. The optimization process may affect the coding region of the auxiliary genes (*e.g.*, vpu, tat and rev genes of HIV-1), and may disrupt RNA secondary structure. Thus, the ancestral sequences can be semi-optimized. A semi-optimized sequence has the optimized sequence for portions of the sequence that do not span other features, where the non-optimized ancestral sequence is used instead. For example, for HIV-1 ancestral sequences, the optimized ancestral sequence is used for portions of the sequence that do not span the vpu, tat, rev and RRE regions, while the "non-optimized" ancestral sequence is used for the portions of the sequence that overlap the vpu, tat, rev and RRE regions.

[0124] *Determination of HIV Ancestral Viral Sequences*

[0125] Ancestral viral sequences can be determined for any gene or genes from HIV type 1 (HIV-1), HIV type 2 (HIV-2), or other HIV viruses, including, for example, for an HIV-1 subtype, for an HIV-2 subtype, for other HIV subtypes, for an emerging HIV subtype, and for HIV variants, such as widely dispersed or geographically isolated variants. For example, an ancestral viral gene sequence can be determined for *env* and *gag* genes of HIV-1, such as for HIV-1 subtypes A, B, C, D, E, F, G, H, J, AG, AGI, and for groups M, N, O, or for HIV-2 viruses or HIV-2 subtypes A or B. In specific embodiments, ancestral viral sequences are determined for *env* genes of HIV-1 subtypes B and/or C, or for *gag* genes from subtypes B and/or C. In other embodiments, the ancestral viral sequence is determined for other HIV genes or polypeptides, such as *nef*, *pol*, or other auxiliary genes or polypeptides.

[0126] Nucleic acid sequences of a selected HIV-1 or HIV-2 gene from presently and/or formerly circulating viruses can be identified from existing databases (*e.g.*, from GenBank or Los Alamos sequence databases). The sequence of circulating viruses can also be determined by recombinant DNA methodologies. (*See, e.g.*, Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual*, 2nd ed., Cold Spring Harbor Publish., Cold Spring Harbor, N.Y. (1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual*, W.H. Freeman, N.Y. (1990); Ausubel *et al.*, *supra.*) For each data set, several sequences from a different viral strain, subtype or group are used as an outgroup to root the sequences of interest. A model of

sequence substitutions and then a maximum likelihood phylogeny is determined for each data set (*e.g.*, subtype and outgroup). The ancestral viral sequence is determined as the sequence at the basal node of the variant sequences (*see, e.g.*, Figures 1 and 2). This ancestral viral sequence is thus approximately equidistant from the different sequences within the subtype.

[0127] In one embodiment, an ancestral HIV-1 group M, subtype B, *env* sequence was determined using 41 distinct isolates. (The determined nucleic acid and amino acid sequences are depicted in Tables 1 and 2 (SEQ ID NO:1 and SEQ ID NO:2), respectively). Referring to Figure 2, 38 subtype B sequences and 3 subtype D (outgroup) sequences were used to root the subtype B sequences. The subtype B sequences were from nine countries, representing a broad sample of subtype B diversity: Australia, 8 sequences; China, 1 sequence; France, 5 sequences; Gabon, 1 sequence; Germany, 2 sequences; Great Britain, 2 sequences; the Netherlands, 2 sequences; Spain, 1 sequence; U.S.A., 15 sequences. The determined ancestor protein is 884 amino acids in length. The distances between this ancestral viral sequence and circulating strains used to determine it were on average 12.3% (range: 8.0-21.0%) while the available specimens were 17.3% different from each other (range: 13.3-23.2%). The ancestor sequence is therefore, on average, more closely related to any given circulating virus than to any other variant. When compared with other subtype B strains, the ancestral sequence is most similar to USAD8 (Theodore *et al.*, *AIDS Res. Human Retrovir.* 12:191-94 (1996)), with an identity of 94.6% at the amino acid level.

[0128] Surprisingly, the determined ancestral viral sequence of the HIV-1 subtype B *env* gene encodes a wide variety of immunologically active peptides when processed for antigen presentation. Nearly all known subtype B CTL epitope consensus amino acids (387/390; 99.23%) are represented in the determined ancestral viral sequence for the subtype B, gp160 sequence. In contrast, most other variants of HIV-1 subtype B have below 95% epitope sequence conservation (although this is a not a necessary feature of ancestral viral sequences, but is a consequence of the rapid expansion of HIV-1). Thus, an immunogenic composition to this subtype B ancestor protein will elicit broad neutralizing antibody against HIV-1 isolates of the same subtype. An immunogenic composition to this subtype B ancestor protein will also elicit a broad cellular response mediated by antigen-specific T-cells.

[0129] In another embodiment, similar computational methods were used to determine the ancestral viral sequence of the HIV-1 subtype C *env* gene sequence. HIV-1 subtype C is widespread in developing countries. Subtype C is the most common subtype worldwide,

responsible for an estimated 30% of HIV-1 infections, and a major component of epidemics in Africa, India and China. The ancestral viral sequence for HIV-1 group M, subtype C, *env* gene was determined using 57 distinct isolates (39 subtype C sequences and 18 outgroup sequences (two from each of the other group M subtypes); Figure 8). The determined amino acid sequence is depicted in Table 4 (SEQ ID NO:4). The determined nucleic acid sequence, optimized for expression in human cells, is depicted in Table 3 (SEQ ID NO:3).

[0130] The subtype C sequences were from twelve African and Asian countries, representing a broad sample of subtype C diversity worldwide: Botswana, 8 sequences; Brazil, 2 sequences; Burundi, 8 sequences; Peoples Republic of China, 1 sequence; Djibouti, 2 sequences; Ethiopia, 1 sequence; India, 8 sequences; Malawi, 3 sequences; Senegal, 1 sequence; Somalia, 1 sequence; Uganda, 1 sequence; and Zambia, 3 sequences. The determined ancestor protein is 853 amino acids in length. The distances between this ancestral viral sequence and circulating strains used to determine it were on average 11.7% (range: 9.3-14.3%) while the available specimens were on average 16.6% different from each other (range: 7.1-21.7%). The ancestor protein sequence is therefore, on average, more closely related to any given circulating virus than to any other variant. When compared with other subtype C strains, the ancestral sequence is most similar to MW965 (Gao *et al.*, *J Virol.* 70:1651-67 (1996)), with an identity of 89.5% at the amino acid level.

[0131] Surprisingly, the determined ancestral viral sequence encodes a wide variety of immunologically active peptides when processed for antigen presentation. Nearly all known subtype C CTL epitope consensus sequences (389/396; 98.23%) are represented in the determined ancestral viral sequence for the subtype C, gp160 sequence. In contrast, typical variants of HIV-1 subtype C (those used to determine the ancestral sequence) have less than 95.19% epitope sequence conservation (average 90.36%, range 64.56 – 95.19%). Thus, a vaccine to this subtype C ancestral viral sequence will elicit broad neutralizing antibody against HIV-1 isolates of the same subtype. An immunogenic composition to this subtype C ancestor protein will also elicit a broad cellular response mediated by antigen-specific T-cells.

[0132] Optimized and semi-optimized sequences for an HIV ancestral sequence are also provided. Ancestral viral sequences can be optimized for expression in particular host cells. While the optimized ancestral sequence encodes the same amino acid sequence for a gene as the non-optimized sequence, the optimized sequence may not be fully functional in a

synthetic virus due to the disruption of auxiliary genes in different reading frames, disruption of the RNA secondary structure, and the like. For example, optimization of the HIV-1 *env* sequence can disrupt the auxiliary genes for *vpu*, *tat* and/or *rev*, and/or the RNA secondary structure Rev responsive element (RRE). Semi-optimized sequences are prepared by using optimized sequences for portions of the sequence that do not span other genes, RNA secondary structure, and the like. For portions of the sequence that overlap such features, the "non-optimized" ancestral sequence is used (*e.g.*, for regions overlapping *vpu*, *tat*, *rev* and/or RRE). In specific embodiments, semi-optimized ancestral viral sequences for HIV-1 subtypes B and C are provided. (*See* Tables 5 (SEQ ID NO: 5) and 6 (SEQ ID NO:6).)

[0133] In other embodiments, ancestral viral sequences are determined for widely circulating variants or geographically-restricted variants. For example, samples can be collected of an HIV-1 subtype which is widely spread (*e.g.* present in many countries or in regions without obvious geographic boundaries). Similarly, samples can be collected of an HIV-1 subtype which is geographically restricted (*e.g.*, to a country, regions or other physically defined area). The sequences of the genes (*e.g.*, *gag* or *env*) in the samples are determined by recombinant DNA methods (*see*, *e.g.*, Sambrook *et al.*, *supra*; Kriegler, *supra*; Ausubel *et al.*, *supra*), or from information in databases. Typically, the number of samples will range from about 20 to about 50, depending on their current availability and the time the virus has been circulating in the region of interest (*e.g.*, the longer the time the virus has been circulating, the greater the diversity and the greater the information to be gleaned from the samples). The ancestral viral sequence, either nucleic acid or amino acid, is then determined using the computational methods described herein.

[0134] *Ancestor Proteins*

[0135] The invention further relates to ancestor proteins based on a determined ancestral viral sequence. Such ancestor proteins include, for example, full-length protein, polypeptides, fragments, derivatives and analogs thereof. In one aspect, the invention provides amino acid sequences of ancestor proteins (*see*, *e.g.*, Tables 2 and 4; SEQ ID NO:2; SEQ ID NO:4). In certain embodiments, the ancestor protein is functionally active. Ancestor proteins, fragments, derivatives and analogs typically have the desired immunogenicity or antigenicity and can be used, for example, in immunoassays, for immunization, in vaccines, and the like. A specific embodiment relates to an ancestor protein, fragment, derivative or analog that can be bound by an antibody. Such ancestor proteins, fragments, derivatives or

analogs can be tested for the desired immunogenicity by procedures known in the art. (*See e.g.*, Harlow and Lane, *supra*).

[0136] In another aspect, a polypeptide is provided which consists of or comprises a fragment that has at least 8-10 contiguous amino acids of the ancestor protein. In other embodiments, the fragment comprises at least 20 or 50 contiguous amino acids of the ancestor protein. In other embodiments, the fragments are not larger than 35, 100 or 200 amino acids.

[0137] Ancestor protein derivatives and analogs can be produced by various methods known in the art. The manipulations which result in their production can occur at the gene or protein level. For example, a nucleic acid encoding an ancestor protein can be modified by any of numerous strategies known in the art (*see, e.g.*, Sambrook *et al.*, *supra*), such as by making conservative substitutions, deletions, insertions, and the like. The nucleic acid sequence can be cleaved at appropriate sites with restriction endonuclease(s), followed by further enzymatic modification, if desired, isolated, and ligated *in vitro*. In the production of nucleic acids encoding a fragment, derivative or analog of an ancestor protein, the modified nucleic acid typically remains in the proper translational reading frame, so that the reading frame is not interrupted by translational stop signals or other signals that interfere with the synthesis of the fragment, derivative or analog. The ancestral viral sequence nucleic acid can also be mutated *in vitro* or *in vivo* to create and/or destroy translation, initiation and/or termination sequences. The ancestral viral sequence-encoding nucleic acid can also be mutated to create variations in coding regions and/or to form new restriction endonuclease sites or destroy preexisting ones and to facilitate further *in vitro* modification. Any technique for mutagenesis known in the art can be used, including but not limited to chemical mutagenesis, *in vitro* site-directed mutagenesis, and the like.

[0138] Manipulations of the ancestral viral sequence can also be made at the protein level. Included within the scope of the invention are ancestor protein fragments, derivatives or analogs that are differentially modified during or after synthesis (*e.g.*, *in vivo* or *in vitro* translation). Such modifications include conservative substitution, glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, linkage to an antibody molecule or other cellular ligand, and the like. Any of numerous chemical modifications can be carried out by known techniques, including, but not limited to, specific chemical cleavage (*e.g.*, by cyanogen bromide); enzymatic cleavage (*e.g.*,

by trypsin, chymotrypsin, papain, V8 protease, and the like); modification by, for example, $NaBH_4$ acetylation, formylation, oxidation and reduction; metabolic synthesis in the presence of tunicamycin; and the like.

[0139] In addition, fragments, derivatives and analogs of ancestor proteins can be chemically synthesized. For example, a peptide corresponding to a portion, or fragment, of an ancestor protein, which comprises a desired domain, can be synthesized by use of chemical synthetic methods using, for example, an automated peptide synthesizer. (*See also* Hunkapiller *et al.*, *Nature* 310:105-11 (1984); Stewart and Young, *Solid Phase Peptide Synthesis*, 2nd ed., Pierce Chemical Co., Rockford, IL, (1984).) Furthermore, if desired, nonclassical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the polypeptide sequence. Non-classical amino acids include, but are not limited to, the D-isomers of the common amino acids, α-amino isobutyric acid, 4-aminobutyric acid, 2-amino butyric acid, 6-amino hexanoic acid, 2-amino isobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxyproline, sarcosine, citrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, β-alanine, selenocysteine, fluoro-amino acids, designer amino acids such as β-methyl amino acids, C α-methyl amino acids, N α-methyl amino acids, and other amino acid analogs. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

[0140] The ancestor protein, fragment, derivative or analog can also be a chimeric, or fusion, protein comprising an ancestor protein, fragment, derivative or analog thereof (typically consisting of at least a domain or motif of the ancestor protein, or at least 10 contiguous amino acids of the ancestor protein) joined at its amino- or carboxy-terminus via a peptide bond to an amino acid sequence of a different protein. In one embodiment, such a chimeric protein is produced by recombinant expression of nucleic acid encoding the chimeric protein. The chimeric nucleic acid can be made by ligating the appropriate nucleic acid sequences to each other in the proper reading frame and expressing the chimeric product by methods commonly known in the art. Alternatively, the chimeric protein can be made by protein synthetic techniques (*e.g.*, by use of an automated peptide synthesizer).

[0141] Ancestor protein can be isolated and purified by standard methods including chromatography (*e.g.*, ion exchange, affinity, sizing column chromatography, high pressure liquid chromatography), centrifugation, differential solubility, or by any other standard technique for the purification of proteins.

[0142] *Determination of COT Sequences*

[0143] The present invention also provides methods for determining a nucleic acid and protein sequences using Center of Tree (or COT) analysis.

[0144] COT provides a method for identifying a position at a node or on a branch of a phylogenetic tree having completely specified branch lengths. This position, called the "center of tree" or COT, is the point at which a specified function F of the lengths from any point to all tips of the tree is minimized. By way of explanation, suppose a tree **T** has n tips or leaves, labelled a_1, a_2,..., a_n, and p is a point on a branch or is a node of the tree. Let l_i be the distance along the tree branches from p to a_i. Then a COT of **T** for the function F, is a point $\hat{p}$ satisfying the following relationship:

$$F(\hat{p}:\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_n) \le F(p:l_1,l_2,\cdots,l_n), \text{ for all points } p,$$

where the notation $F(p:l_1,l_2,\cdots,l_n)$ highlights the fact that the distances l_i depend on the point p.

[0145] In this description, the form of F is general; specific choices for F can be made, based on the intended application. A general algorithm that is applicable for most useful continuous Fs is described (*infra*). For a given choice of F, an algorithm based on COT analysis can be selected. A COT-based algorithm can be selected, for example, to be more efficient. Such an algorithm is described to find COT when F is the mean of squares (MS) of the ls. Depending upon F, one or more COTs may exist for a given tree, but for many reasonable choices of F, the COT will be unique.

[0146] *General Algorithm*

[0147] First, for a certain large class of functions $F: R^n \to R$, namely those which are continuous and have finitely many extreme points (which includes those functions described *infra*), that there is a finite number of points along the tree which are possible COTs, that can be enumerated constructively, and determine which points are in fact COTs.

[0148] For an unrooted tree **T** of n tips, there are $u \le 2n-2$ nodes, counting tips and internal branches, and $w \le 2n-3$ branches, including internal and external branches. (u and w are less than their maxima when polytomies exist in the tree.) For each node $q_j, 1 \le j \le u$, $c_j = F(q_j:l_1,l_2,\cdots,l_n)$ is calculated. Each q_j is a candidate COT.

[0149] The branches are enumerated $b_k, 1 \le k \le w$. Candidate COTs are determined for each branch. Note each branch, say branch k, is flanked by two nodes, call them R_k and L_k (suggesting right and left nodes). Let the branch length of b_k be l. Now the tree is divided into two parts, call them right and left, so that if the tree had a root within branch k, the tips $a_1, a_2, \ldots, a_n$ would be divided into two groups, those descended from R_k and those descended from L_k. Suppose there are s right tips and t left tips. Let the distances from the right tips to R_k be written $\rho_1, \cdots \rho_s$, and the distances from the left tips to L_k be written $\lambda_1, \cdots \lambda_t$. Now, let a point p lying on branch b_k be a distance x from the right node R_k. Then the distance from p to L_k is $l - x$.

[0150] Then for branch b_k and p defined along it as described above,

$$F(p : l_1, \cdots, l_n) = F(p : \rho_1 + x, \cdots, \rho_s + x, \lambda_1 + l - x, \cdots, \lambda_t + l - x) = \tilde{F}_{b_k}(x),\ 0 < x < l.$$

In other words, on any branch k of **T**, the function F of n distances can be expressed for every point p along that branch as a function of a single variable x. By this assumption that F has a finite number of extreme points, the functions $\tilde{F}_{b_k}(x)$ have a finite number of minima for x between 0 and l. Because F is continuous, those minima can be found by standard numerical methods, and each minimum $\hat{x}$ is associated with a point $p_{\hat{x}}$ as described above. Suppose there are v such points over all w branches (v may be greater than, equal to, or less than w). Then the following equation can be written: $d_i = F(p_i : l_1, \cdots, l_n) = \tilde{F}_{b(p_i)}(\hat{x}_i)$, for $1 \le i \le v$, where $b(p_i)$ is the branch associated with point p_i (not necessarily the ith branch). Then each p_i is a candidate COT, since if p_i is to minimize F among all points on the tree, it must at least minimize F on those points comprising the branch on which p resides. Since the nodes and branches contain all points on the tree, all possible COTs in the q_j and the p_i have been enumerated.

[0151] Therefore, according to the definition of a COT in the first paragraph, any point $p \in \{q_1, \cdots, q_u, p_1, \cdots, p_v\}$ is a COT that satisfies

$$F(p : l_1, \cdots, l_n) = \min\{c_1, \cdots, c_u, d_1, \cdots, d_v\}$$

and all such points p are the only COTs for tree **T** given function F.

[0152] This decomposition of possible COTs into separate consideration of nodes and branches allows phylogenetic trees to be expressed in computer programs as data structures

that can be efficiently traversed by recursive routines which isolate each node and branch individually and systematically. The above decomposition formally describes the tasks to be performed upon consideration of each node and branch. During the algorithm, the points and function values are stored, and the final determination of COTs is easily accomplished by identifying the minima of the list of values and their associated points after the tree data structure has been completely traversed.

[0153] *Algorithm to find points minimizing the mean squared distance from the points to tips (least-squares or LS-COT)*

[0154] In this special case, let $F(p: l_1,\cdots,l_n)=\frac{1}{n}\sum_{i=1}^{n} l_i^2$, the mean of the squared distances from the tips to point p. The COT obtained by minimizing this function essentially balances the average length of the branches on either side of point p, and as such provides a point which will yield a single reconstructed sequence that has the greatest amount of sequence similarity to all the tips as possible, given the evolutionary constraints of nucleotide change along the tree branches. As in the general algorithm, the tree is decomposed into nodes and branches, all possible COTs are enumerated, and F is calculated for each possibility. The point with the minimum F is the COT. The function F can be expressed in terms of quantities that can be efficiently calculated as the tree is traversed recursively; this allows the algorithm to accumulate the quantities c_i and d_i. First, the method of identifying possible COTs and calculating c_i and d_i based on these quantities is described; then the recursion equations for these quantities is described that can be using in the tree-traversal algorithm.

[0155] *Nodes*: Consider each node q_i as a temporary root of the tree, and suppose q_i has k descendant branches, each of which defines a subtree with t_m tips, $1 \le m \le k$. Then F can be written

$$
\begin{aligned}
c_i &= \frac{1}{n}\sum_{j=1}^{n} l_j^2 \\
&= \gamma_1\left(\frac{1}{t_1}\sum_{j=1}^{t_1}(l_j^{(1)})^2\right)+\cdots+\gamma_k\left(\frac{1}{t_k}\sum_{j=1}^{t_k}(l_j^{(k)})^2\right) \\
&= \gamma_1 MS_1+\cdots+\gamma_k MS_k,
\end{aligned}
$$

where $\gamma_m=\frac{t_m}{n}$, and $(l_j^{(m)}), 1 \le j \le t_m$, are the distances from q_i to each of the t_m tips of the mth subtree.

[0156] Each MS_m is therefore the mean of squared distances of the branches of subtree m, considering node q_i as the root, and each γ_m is the proportion of the n tips of the entire tree associated with subtree m.

[0157] *Branches*: With this function F, there exists at most one possible COT on any branch. Consider a branch of length l with left and right nodes as described in the general algorithm above, and consider a point p within the branch. Let M_L be the simple average of distances from point p to the left tips, and M_R be the average of distances from p to the right tips. Suppose there are t left tips and s right tips, and let $\gamma = \frac{t}{n}$. Now, define α as follows:

$$\alpha = \frac{(1-\gamma)M_R - \gamma M_L}{l} + 1 - \gamma .$$

[0158] Then there is a possible COT within the branch if $0 < \alpha < 1$, and it is the distance αl from the *left* node along the branch. If there is such a point, then the value of F at that point, d_i, can be written as

$$d_i = \gamma(1-\gamma)(M_L + M_R + l) - \gamma(M_L^2 - MS_L) - (1-\gamma)(M_R^2 - MS_R),$$

where MS is the mean of summed squared distances from the left or right nodes to their descendant tips as indicated.

[0159] Finally, the COT is the point which is associated with the smallest value among the c_i and d_i.

[0160] *Algorithm to find points minimizing the mean distance from the points to tips ("Minimum of means" or MM-COT)*

[0161] Another useful and simpler function is $F(p : l_1, \cdots, l_n) = \frac{1}{n}\sum_{i=1}^{n} l_i = \frac{M_L + M_R}{2}$ for points within branches, and M_q for nodes. For points within branches, the condition that must be met such that p is a possible COT is the inequality involving α above, with γ set equal to 0.5. The "minimum of means" COT is then $\min\{c_1, \cdots, c_u, d_1, \cdots, d_v\}$, where now $c_i = M_{q(i)}$ for each node i, and $d_i = \frac{M_{L(i)} + M_{R(i)}}{2}$ for points on branches.

[0162] *Recursions to calculate M and MS for the above algorithms*

[0163] Suppose node q has k descendant nodes. Each of the k nodes is connected to q by a branch of length l_i, and each is the root of a subtree having s_i tips, $1 \le i \le k$. Suppose for each subtree, the mean distance M_i and the mean squared distances MS_i from node to tips have been calculated. Then the mean distance M_q and mean squared distance MS_q from q to all $s = s_1 + \cdots + s_k$ descendant tips are given by:

$$M_q = \gamma_1(M_1 + l_1) + \cdots + \gamma_k(M_k + l_k), \text{ and}$$
$$MS_q = \gamma_1(MS_1 + 2l_1M_1 + l_1^2) + \cdots + \gamma_k(MS_k + 2l_kM_k + l_k^2),$$
where $\gamma_i = \frac{s_i}{s}$ for $1 \le i \le k$.

[0164] These quantities can thus be built up as a tree is recursively traversed, and can be used in the calculations described above.

[0165] As will be appreciated by the skilled, the methodology for COT analysis can be applied to determine the nucleic acid sequences for a highly disperse viruses. In addition to the embodiments of the determination of HIV COT nucleic acid and protein sequences (*infra*), COT analysis can be used to determine nucleic acid and protein sequences for other highly disperse viruses.

[0166] *HIV COT Sequences*

[0167] COT viral sequences can be determined for any gene or genes from HIV type 1 (HIV-1), HIV type 2 (HIV-2), or other HIV viruses, including, for example, for an HIV-1 subtype, for an HIV-2 subtype, for other HIV subtypes, for an emerging HIV subtype, and for HIV variants, such as widely dispersed or geographically isolated variants. For example, an COT viral gene sequence can be determined for *env* and *gag* genes of HIV-1, such as for HIV-1 subtypes A, B, C, D, E, F, G, H, J, AG, AGI, and for groups M, N, O, or for HIV-2 viruses or HIV-2 subtypes A or B. In specific embodiments, COT viral sequences are determined for *env* genes of HIV-1 subtypes B and/or C, or for *gag* genes from subtypes B and/or C. In other embodiments, the COT viral sequence is determined for other HIV genes or polypeptides, such as *nef*, *pol*, or other auxiliary genes or polypeptides.

[0168] Nucleic acid sequences of a selected HIV-1 or HIV-2 gene from presently and/or formerly circulating viruses can be identified from existing databases (*e.g.*, from GenBank or

Los Alamos sequence databases). The sequence of circulating viruses can also be determined by recombinant DNA methodologies. (*See, e.g.*, Sambrook *et al.*, *Molecular Cloning, A Laboratory Manual*, 2nd ed., Cold Spring Harbor Publish., Cold Spring Harbor, N.Y. (1989); Kriegler, *Gene Transfer and Expression: A Laboratory Manual*, W.H. Freeman, N.Y. (1990); Ausubel *et al.*, *supra.*) For each data set, several sequences from a different viral strain, subtype or group are used as an outgroup to root the sequences of interest.

[0169] The determined COT viral gene sequence is more closely related, on average, to a gene sequence of any given circulating virus than to any other variant. In certain embodiments, the COT viral gene sequence has at least 70% identity with the LScot and MMcot sequences set forth in Figures 9 to 17 or 27-35, but does not have 100% identity with any circulating viral variant. In specific embodiments, the COT viral gene sequence is an LScot or MMcot sequence set forth in Figures 9 to 17, but does not have 100% identity with any circulating viral variant. In additional specific embodiments, the COT viral gene sequence is an LScot or MMcot sequence set forth in Figures 27 to 35, but does not have 100% identity with any circulating viral variant.

[0170] Optimized and semi-optimized sequences for an HIV COT sequence are also provided. COT viral sequences can be optimized for expression in particular host cells. While the optimized COT sequence encodes the same amino acid sequence for a gene as the non-optimized sequence, the optimized sequence may not be fully functional in a synthetic virus due to the disruption of auxiliary genes in different reading frames, disruption of the RNA secondary structure, and the like. For example, optimization of the HIV-1 *env* sequence can disrupt the auxiliary genes for *vpu*, *tat* and/or *rev*, and/or the RNA secondary structure Rev responsive element (RRE). Semi-optimized sequences are prepared by using optimized sequences for portions of the sequence that do not span other genes, RNA secondary structure, and the like. For portions of the sequence that overlap such features, the "non-optimized" COT sequence is used (*e.g.*, for regions overlapping *vpu*, *tat*, *rev* and/or RRE). In specific embodiments, semi-optimized COT viral sequences for HIV-1 subtypes B and C are provided.

[0171] In other embodiments, COT viral sequences are determined for widely circulating variants or geographically-restricted variants. For example, samples can be collected of an HIV-1 subtype which is widely spread (*e.g.* present in many countries or in regions without obvious geographic boundaries). Similarly, samples can be collected of an HIV-1 subtype

which is geographically restricted (*e.g.*, to a country, regions or other physically defined area). The sequences of the genes (*e.g.*, *gag* or *env*) in the samples are determined by recombinant DNA methods (*see*, *e.g.*, Sambrook *et al.*, *supra*; Kriegler, *supra*; Ausubel *et al.*, *supra*), or from information in databases. Typically, the number of samples will range from about 20 to about 50, depending on their current availability and the time the virus has been circulating in the region of interest (*e.g.*, the longer the time the virus has been circulating, the greater the diversity and the greater the information to be gleaned from the samples). The COT viral sequence, either nucleic acid or amino acid, is then determined using the computational methods described herein.

[0172] *COT Proteins*

[0173] The invention further relates to COT proteins based on a determined COT viral sequence. Such COT proteins include, for example, full-length protein, polypeptides, fragments, derivatives and analogs thereof. In one aspect, amino acid sequences of COT proteins are provided. In certain embodiments, the COT protein is functionally active. COT proteins, fragments, derivatives and analogs typically are immunogenic or antigenic and can be used, for example, in immunoassays, for immunization, in vaccines, and the like. A specific embodiment relates to an COT protein, fragment, derivative or analog that can be bound by an antibody. Such COT proteins, fragments, derivatives or analogs can be tested for the desired immunogenicity by procedures known in the art. (*See e.g.*, Harlow and Lane, *supra*). In specific embodiments, an isolated COT protein has the sequence of an LScot or MMcot amino acid sequence set forth in Figures 27 to 25 and 36 to 44.

[0174] In another aspect, a polypeptide is provided which consists of or comprises a fragment that has at least 8-10 contiguous amino acids of the COT protein. In other embodiments, the fragment comprises at least 20 or 50 contiguous amino acids of the COT protein. In other embodiments, the fragments are not larger than 35, 100 or 200 amino acids. In specific embodiments, an isolated antigenic COT protein fragment is an antigenic fragment of an LScot or MMcot amino acid sequence set forth in Figures 27 to 25 and 36 to 44.

[0175] COT protein derivatives and analogs can be produced by various methods known in the art. The manipulations which result in their production can occur at the gene or protein level. For example, a nucleic acid encoding a COT protein can be modified by any of numerous strategies known in the art (*see*, *e.g.*, Sambrook *et al.*, *supra*), such as by making

conservative substitutions, deletions, insertions, and the like. The nucleic acid sequence can be cleaved at appropriate sites with restriction endonuclease(s), followed by further enzymatic modification, if desired, isolated, and ligated *in vitro*. In the production of nucleic acids encoding a fragment, derivative or analog of a COT protein, the modified nucleic acid typically remains in the proper translational reading frame, so that the reading frame is not interrupted by translational stop signals or other signals that interfere with the synthesis of the fragment, derivative or analog. The COT viral sequence nucleic acid can also be mutated *in vitro* or *in vivo* to create and/or destroy translation, initiation and/or termination sequences. The COT viral sequence-encoding nucleic acid can also be mutated to create variations in coding regions and/or to form new restriction endonuclease sites or destroy preexisting ones and to facilitate further *in vitro* modification. Any technique for mutagenesis known in the art can be used, including but not limited to chemical mutagenesis, *in vitro* site-directed mutagenesis, and the like.

[0176] Manipulations of the COT viral sequence can also be made at the protein level. Included within the scope of the invention are COT protein fragments, derivatives or analogs that are differentially modified during or after synthesis (*e.g.*, *in vivo* or *in vitro* translation). Such modifications include conservative substitution, glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, linkage to an antibody molecule or other cellular ligand, and the like. Any of numerous chemical modifications can be carried out by known techniques, including, but not limited to, specific chemical cleavage (*e.g.*, by cyanogen bromide); enzymatic cleavage (*e.g.*, by trypsin, chymotrypsin, papain, V8 protease, and the like); modification by, for example, $NaBH_4$ acetylation, formylation, oxidation and reduction; metabolic synthesis in the presence of tunicamycin; and the like.

[0177] In addition, fragments, derivatives and analogs of COT proteins can be chemically synthesized. For example, a peptide corresponding to a portion, or fragment, of an COT protein, which comprises a desired domain, can be synthesized by use of chemical synthetic methods using, for example, an automated peptide synthesizer. (*See also* Hunkapiller *et al.*, *Nature* 310:105-11 (1984); Stewart and Young, *Solid Phase Peptide Synthesis*, 2nd ed., Pierce Chemical Co., Rockford, IL, (1984).) Furthermore, if desired, nonclassical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the polypeptide sequence. Non-classical amino acids include, but are not limited to, the D-isomers of the common amino acids, α-amino isobutyric acid, 4-aminobutyric acid, 2-amino

butyric acid, 6-amino hexanoic acid, 2-amino isobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxyproline, sarcosine, citrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, β-alanine, selenocysteine, fluoro-amino acids, designer amino acids such as β-methyl amino acids, C α-methyl amino acids, N α-methyl amino acids, and other amino acid analogs. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

[0178] The COT protein, fragment, derivative or analog can also be a chimeric, or fusion, protein comprising an COT protein, fragment, derivative or analog thereof (typically consisting of at least a domain or motif of the COT protein, or at least 10 contiguous amino acids of the COT protein) joined at its amino- or carboxy-terminus via a peptide bond to an amino acid sequence of a different protein. In one embodiment, such a chimeric protein is produced by recombinant expression of nucleic acid encoding the chimeric protein. The chimeric nucleic acid can be made by ligating the appropriate nucleic acid sequences to each other in the proper reading frame and expressing the chimeric product by methods commonly known in the art. Alternatively, the chimeric protein can be made by protein synthetic techniques (*e.g.*, by use of an automated peptide synthesizer).

[0179] COT protein can be isolated and purified by standard methods including chromatography (*e.g.*, ion exchange, affinity, sizing column chromatography, high pressure liquid chromatography), centrifugation, differential solubility, or by any other standard technique for the purification of proteins.

[0180] *Nucleic Acids Encoding Ancestral or COT Viral Sequences*

[0181] Once an ancestral or COT viral sequence is determined by the methods described herein, recombinant DNA methods can be used to prepare nucleic acids encoding the ancestral or COT viral sequence of interest. Suitable methods include, but are not limited to: (1) modifying an existing viral strain most similar to the ancestor viral sequence; (2) synthesizing a nucleic acid encoding the ancestral or COT viral sequence by joining shorter oligonucleotides (*e.g.*, 160-200 nucleotides in length); or (3) a combination of these methods (*e.g.*, by modifying an existing sequence using fragments with very high similarity to the ancestral or COT viral sequence, while synthesizing *de novo* more divergent sequences).

[0182] The nucleic acid sequences can be produced and manipulated using routine techniques. (*See*, *e.g.*, Sambrook *et al.*, *supra*; Kriegler, *supra*; Ausubel *et al.*, *supra*.)

Table 1 (SEQ ID NO:1)

```
   1  ATGCGCGTGA AGGGCATCCG CAAGAACTAC CAGCACCTGT GGCGCTGGGG
  51  CACCATGCTG CTGGGGATGC TGATGATCTG CTCCGCGGCC GAGAAGCTGT
 101  GGGTGACCGT GTACTACGGC GTGCCCGTGT GGAAGGAGGC CACCACCACC
 151  CTGTTCTGCG CCAGCGACGC CAAGGCTTAC GACACCGAGG TCCACAACGT
 201  GTGGGCCACC CACGCCTGCG TGCCCACCGA CCCCAACCCC CAGGAGGTGG
 251  TGCTGGAGAA CGTGACCGAG AACTTCAACA TGTGGAAGAA CAACATGGTG
 301  GAGCAGATGC ACGAGGACAT CATCAGCCTG TGGGACCAGA GCCTGAAGCC
 351  CTGCGTGAAG TTAACCCCCC TGTGCGTGAC CCTGAACTGC ACCGACGACC
 401  TGCGCACCAA CGCCACCAAC ACCACCAACA GCAGCGCCAC CACCAACACC
 451  ACCAGCAGCG GCGGCGGCAC GATGGAGGGC GAGAAGGGCG AGATCAAGAA
 501  CTGCAGCTTC AACGTGACCA CCAGCATCCG CGACAAGATG CAGAAGGAGT
 551  ACGCCCTGTT CTACAAGCTG GACGTGGTGC CCATCGACAA CGACAACAAC
 601  AACACCAACA ACAACACCAG CTACCGCCTC ATCAACTGCA ACACCAGCGT
 651  GATCACCCAG GCCTGCCCCA AGGTGAGCTT CGAGCCCATC CCCATCCACT
 701  ACTGCACCCC CGCCGGCTTC GCCATCCTGA AGTGCAACGA CAAGAAGTTC
 751  AACGGCACCG GCCCCTGCAC CAACGTGAGC ACCGTGCAGT GCACCCACGG
 801  CATCCGCCCC GTGGTGAGCA CCCAGCTGCT GCTGAACGGC AGCCTGGCCG
 851  AGGAGGAGGT GGTGATCCGC AGCGAGAACT TCACCGACAA CGCCAAGACC
 901  ATCATCGTGC AGCTGAACGA GAGCGTGGAG ATCAACTGCA CGCGTCCCAA
 951  CAACAACACC CGCAAGAGCA TCCCCATCGG CCCTGGCCGC GCCCTGTACG
1001  CCACCGGCAA GATCATCGGC GACATCCGCC AGGCCCACTG CAACCTGTCG
1051  CGAGCCAAGT GGAACAACAC CCTGAAGCAG ATCGTGACCA AGCTGCGCGA
1101  GCAGTTCGGC AACAACAAGA CCACCATCGT GTTCAACCAG AGCAGCGGCG
1151  GCGACCCCGA GATCGTGATG CACAGCTTCA ACTGCGGCGG CGAATTCTTC
1201  TACTGCAACA GCACCCAGCT GTTCAACAGC ACCTGGCACT TCAACGGCAC
1251  CTGGGGCAAC AACAACACCG AGCGCAGCAA CAACGCCGCC GACGACAACG
1301  ACACCATCAC CCTGCCCTGC CGCATCAAGC AGATCATCAA CATGTGGCAG
1351  GAGGTGGGCA AGGCCATGTA CGCCCCCCCC ATCAGCGGCC AGATCCGCTG
1401  CAGCAGCAAC ATCACCGGCC TGCTGCTGAC TCGAGACGGC GGCAACAACG
1451  AGAACACCAA CAACACCGAC ACCGAGATCT TCCGCCCCGG GGGCGGCGAC
1501  ATGCGCGACA ACTGGCGCAG CGAGCTGTAC AAGTACAAGG TGGTGAAGAT
1551  CGAGCCCCTG GGCGTGGCCC CCACCAAGGC CAAGCGCCGC GTGGTGCAGC
1601  GCGAGAAGCG CGCCGTGGGC ATGCTGGGCG CCATGTTCCT GGGCTTCCTG
1651  GGCGCCGCCG GCAGCACCAT GGGCGCCGCC AGCATGACCC TGACCGTGCA
1701  GGCCCGCCAG CTGCTGAGCG GCATCGTGCA GCAGCAGAAC AACCTGCTGC
1751  GCGCCATCGA GGCCCAGCAG CACCTGCTGC AGCTGACCGT GTGGGGCATC
1801  AAGCAGCTGC AGGCCCGCGT GCTGGCCGTG GAGCGGTACC TGAAGGACCA
1851  GCAGCTGCTG GGCATCTGGG GCTGCAGCGG CAAGCTGATC TGCACCACCG
```

```
1901  CGGTGCCCTG GAACGCCAGC TGGAGCAACA AGAGCCTGGA CAAGATCTGG
1951  AACAACATGA CCTGGATGGA GTGGGAGCGC GAGATCGACA ACTACACCGG
2001  CCTGATCTAC ACCCTGATCG AGGAGAGCCA GAACCAGCAG GAGAAGAACG
2051  AGCAGGAGCT GCTGGAGCTG GACAAGTGGG CCAGCCTGTG GAACTGGTTC
2101  GATATCACCA ACTGGCTGTG GTACATCAAG ATCTTCATCA TGATCGTGGG
2151  CGGCCTGGTG GGCCTGCGCA TCGTGTTCGC CGTGCTGAGC ATCGTGAACC
2201  GCGTGCGCCA GGGCTACAGC CCCCTGAGCT TCCAGACCCG CCTGCCCGCC
2251  CCCCGCGGCC CCGACCGCCC CGAGGGCATC GAGGAGGAGG GCGGCGAGCG
2301  CGACCGCGAC CGCAGCGGGC GCCTGGTGAA CGGCTTCCTG GCCCTGATCT
2351  GGGACGACCT GCGCAGCCTG TGCCTGTTCA GCTACCACCG CCTGCGCGAC
2401  CTGCTGCTGA TCGTGGCCCG CATCGTGGAG CTGCTGGGCC GGCGCGGCTG
2451  GGAGGCCCTG AAGTATTGGT GGAACCTGCT GCAGTACTGG AGCCAGGAGC
2501  TGAAGAACAG CGCCGTGAGC CTGCTGAACG CCACCGCCAT CGCCGTGGCC
2551  GAGGGCACCG ACCGCGTGAT CGAGGTGGTG CAGCGCGCCT GCCGCGCCAT
2601  CCTGCACATC CCCCGCCGCA TCCGCCAGGG CCTGGAGCGC GCCCTGCTGT
2651  GA
```

Table 2 (SEQ ID NO:2)

MRVKGIRKNY QHLWRWGTML LGMLMICSAA EKLWVTVYYG VPVWKEATTT LFCASDAKAY
DTEVHNVWAT HACVPTDPNP QEVVLENVTE NFNMWKNNMV EQMHEDIISL WDQSLKPCVK
LTPLCVTLNC TDDLRTNATN TTNSSATTNT TSSGGGTMEG EKGEIKNCSF NVTTSIRDKM
QKEYALFYKL DVVPIDNDNN NTNNNTSYRL INCNTSVITQ ACPKVSFEPI PIHYCTPAGF
AILKCNDKKF NGTGPCTNVS TVQCTHGIRP VVSTQLLLNG SLAEEEVVIR SENFTDNAKT
IIVQLNESVE INCTRPNNNT RKSIPIGPGR ALYATGKIIG DIRQAHCNLS RAKWNNTLKQ
IVTKLREQFG NNKTTIVFNQ SSGGDPEIVM HSFNCGGEFF YCNSTQLFNS TWHFNGTWGN
NNTERSNNAA DDNDTITLPC RIKQIINMWQ EVGKAMYAPP ISGQIRCSSN ITGLLLTRDG
GNNENTNNTD TEIFRPGGGD MRDNWRSELY KYKVVKIEPL GVAPTKAKRR VVQREKRAVG
MLGAMFLGFL GAAGSTMGAA SMTLTVQARQ LLSGIVQQQN NLLRAIEAQQ HLLQLTVWGI
KQLQARVLAV ERYLKDQQLL GIWGCSGKLI CTTAVPWNAS WSNKSLDKIW NNMTWMEWER
EIDNYTGLIY TLIEESQNQQ EKNEQELLEL DKWASLWNWF DITNWLWYIK IFIMIVGGLV
GLRIVFAVLS IVNRVRQGYS PLSFQTRLPA PRGPDRPEGI EEEGGERDRD RSGRLVNGFL
ALIWDDLRSL CLFSYHRLRD LLLIVARIVE LLGRRGWEAL KYWWNLLQYW SQELKNSAVS
LLNATAIAVA EGTDRVIEVV QRACRAILHI PRRIRQGLER ALL

Table 3 (SEQ ID NO:3)

ATGCGGGTGATGGGCATCCTGCGGAACTGCCAGCAGTGGTGGATCTGGGGCATCCTGGGC
TTCTGGATGCTGATGATCTGCAGCGTGATGGGCAACCTGTGGGTGACCGTGTACTACGGC
GTGCCCGTGTGGAAGGAGGCCAAGACCACCCTGTTCTGCGCCAGCGACGCCAAGGCCTAC
GAGCGGGAGGTGCACAACGTGTGGGCCACCCACGCCTGCGTGCCCACCGACCCCAACCCC
CAGGAGATGGTGCTGGAGAACGTGACCGAGAACTTCAACATGTGGAAGAACGACATGGTG
GACCAGATGCACGAGGACATCATCAGCCTGTGGGACCAGAGCCTGAAGCCCTGCGTGAAG
CTGACCCCCCTGTGCGTGACCCTGAACTGCACCAACGTGACCAACACCAACAACAACAAC
AACACCAGCATGGGCGGCGAGATCAAGAACTGCAGCTTCAACATCACCACCGAGCTGCGG
GACAAGAAGCAGAAGGTGTACGCCCTGTTCTACCGGCTGGACATCGTGCCCCTGAACGAG
AACAGCAACAGCAACAGCAGCGAGTACCGGCTGATCAACTGCAACACCAGCGCCATCACC
CAGGCCTGCCCCAAGGTGAGCTTCGACCCCATCCCCATCCACTACTGCGCCCCCGCCGGC
TACGCCATCCTGAAGTGCAACAACAAGACCTTCAACGGCACCGGCCCCTGCAACAACGTG
AGCACCGTGCAGTGCACCCACGGCATCAAGCCCGTGGTGAGCACCCAGCTGCTGCTGAAC
GGCAGCCTGGCCGAGGAGGAGATCATCATCCGGAGCGAGAACCTGACCAACAACGCCAAG
ACCATCATCGTGCACCTGAACGAGAGCGTGGAGATCGTGTGCACCCGGCCCAACAACAAC
ACCCGGAAGAGCATCCGGATCGGCCCCGGCCAGACCTTCTACGCCACCGGCGACATCATC
GGCGACATCCGGCAGGCCCACTGCAACATCAGCGAGAAGGAGTGGAACAAGACCCTGCAG
CGGGTGGGCAAGAAGCTGAAGGAGCACTTCCCCAACAAGACCATCAAGTTCGAGCCCAGC
AGCGGCGGCGACCTGGAGATCACCACCCACAGCTTCAACTGCCGGGGCGAGTTCTTCTAC
TGCAACACCAGCAAGCTGTTCAACAGCACCTACAACAGCACCAACAACGGCACCACCAGC
AACAGCACCATCACCCTGCCCTGCCGGATCAAGCAGATCATCAACATGTGGCAGGGCGTG
GGCCGGGCCATGTACGCCCCCCCCATCGCCGGCAACATCACCTGCAAGAGCAACATCACC
GGCCTGCTGCTGACCCGGGACGGCGGCAACACCAACAACACCACCGAGACCTTCCGGCCC
GGCGGCGGCGACATGCGGGACAACTGGCGGAGCGAGCTGTACAAGTACAAGGTGGTGGAG
ATCAAGCCCCTGGGCGTGGCCCCCACCGAGGCCAAGCGGCGGGTGGTGGAGCGGGAGAAG
CGGGCCGTGGGCATCGGCGCCGTGTTCCTGGGCTTCCTGGGCGCCGCCGGCAGCACCATG
GGCGCCGCCAGCATCACCCTGACCGTGCAGGCCCGGCAGCTGCTGAGCGGCATCGTGCAG
CAGCAGAGCAACCTGCTGCGGGCCATCGAGGCCCAGCAGCACATGCTGCAGCTGACCGTG
TGGGGCATCAAGCAGCTGCAGACCCGGGTGCTGGCCATCGAGCGGTACCTGAAGGACCAG
CAGCTGCTGGGCATCTGGGGCTGCAGCGGCAAGCTGATCTGCACCACCGCCGTGCCCTGG
AACAGCAGCTGGAGCAACAAGAGCCAGGACGACATCTGGGACAACATGACCTGGATGCAG
TGGGACCGGGAGATCAGCAACTACACCGACACCATCTACCGGCTGCTGGAGGACAGCCAG
AACCAGCAGGAGAAGAACGAGAAGGACCTGCTGGCCCTGGACAGCTGGAAGAACCTGTGG
AACTGGTTCGACATCACCAACTGGCTGTGGTACATCAAGATCTTCATCATGATCGTGGGC
GGCCTGATCGGCCTGCGGATCATCTTCGCCGTGCTGAGCATCGTGAACCGGGTGCGGCAG
GGCTACAGCCCCCTGAGCTTCCAGACCCTGACCCCCAACCCCCGGGGCCCCGACCGGCTG
GGCGGCATCGAGGAGGAGGGCGGCGAGCAGGACCGGGACCGGAGCATCCGGCTGGTGAGC
GGCTTCCTGGCCCTGGCCTGGGACGACCTGCGGAGCCTGTGCCTGTTCAGCTACCACCGG

CTGCGGGACTTCATCCTGATCGCCGCCCGGGGCGTGAACCTGCTGGGCCGGAGCAGCCTG
CGGGGCCTGCAGCGGGGCTGGGAGGCCCTGAAGTACCTGGGCAGCCTGGTGCAGTACTGG
GGCCTGGAGCTGAAGAAGAGCGCCATCAGCCTGCTGGACACCATCGCCATCGCCGTGGCC
GAGGGCACCGACCGGATCATCGAGCTGGTGCAGCGGATCTGCCGGGCCATCCGGAACATC
CCCCGGCGGATCCGGCAGGGCTTCGAGGCCGCCCTGCAGTGA

Table 4 (SEQ ID NO:4)

MRVMGILRNCQQWWIWGILGFWMLMICSVMGNLWVTVYYGVPVWKEAKTT
LFCASDAKAYEREVHNVWATHACVPTDPNPQEMVLENVTENFNMWKNDMV
DQMHEDIISLWDQSLKPCVKLTPLCVTLNCTNVTNTNNNNNTSMGGEIKN
CSFNITTELRDKKQKVYALFYRLDIVPLNENSNSNSSEYRLINCNTSAIT
QACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGTGPCNNVSTVQCTHGIK
PVVSTQLLLNGSLAEEEIIIRSENLTNNAKTIIVHLNESVEIVCTRPNNN
TRKSIRIGPGQTFYATGDIIGDIRQAHCNISEKEWNKTLQRVGKKLKEHF
PNKTIKFEPSSGGDLEITTHSFNCRGEFFYCNTSKLFNSTYNSTNNGTTS
NSTITLPCRIKQIINMWQGVGRAMYAPPIAGNITCKSNITGLLLTRDGGN
TNNTTETFRPGGGDMRDNWRSELYKYKVVEIKPLGVAPTEAKRRVVEREK
RAVGIGAVFLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQSNLLRAIE
AQQHMLQLTVWGIKQLQTRVLAIERYLKDQQLLGIWGCSGKLICTTAVPW
NSSWSNKSQDDIWDNMTWMQWDREISNYTDTIYRLLEDSQNQQEKNEKDL
LALDSWKNLWNWFDITNWLWYIKIFIMIVGGLIGLRIIFAVLSIVNRVRQ
GYSPLSFQTLTPNPRGPDRLGGIEEEGGEQDRDRSIRLVSGFLALAWDDL
RSLCLFSYHRLRDFILIAARGVNLLGRSSLRGLQRGWEALKYLGSLVQYW
GLELKKSAISLLDTIAIAVAEGTDRIIELVQRICRAIRNIPRRIRQGFEA
ALQ

Table 5 (SEQ ID NO:5)

ATGAGAGTGAAGGGGATCAGGAAGAACTATCAGCACTTGTGGAGATGGGG
CACCATGCTCCTTGGGATGTTGATGATCTGTAGCGCCGCCGAGAAGCTGT
GGGTGACCGTGTACTACGGCGTGCCCGTGTGGAAGGAGGCCACCACCACC
CTGTTCTGCGCCAGCGACGCCAAGGCTTACGACACCGAGGTCCACAACGT
GTGGGCCACCCACGCCTGCGTGCCCACCGACCCCAACCCCCAGGAGGTGG
TGCTGGAGAACGTGACCGAGAACTTCAACATGTGGAAGAACAACATGGTG
GAGCAGATGCACGAGGACATCATCAGCCTGTGGGACCAGAGCCTGAAGCC
CTGCGTGAAGTTAACCCCCCTGTGCGTGACCCTGAACTGCACCGACGACC
TGCGCACCAACGCCACCAACACCACCAACAGCAGCGCCACCACCAACACC
ACCAGCAGCGGCGGCGGCACGATGGAGGGCGAGAAGGGCGAGATCAAGAA
CTGCAGCTTCAACGTGACCACCAGCATCCGCGACAAGATGCAGAAGGAGT
ACGCCCTGTTCTACAAGCTGGACGTGGTGCCCATCGACAACGACAACAAC
AACACCAACAACAACACCAGCTACCGCCTCATCAACTGCAACACCAGCGT
GATCACCCAGGCCTGCCCCAAGGTGAGCTTCGAGCCCATCCCCATCCACT
ACTGCACCCCCGCCGGCTTCGCCATCCTGAAGTGCAACGACAAGAAGTTC
AACGGCACCGGCCCCTGCACCAACGTGAGCACCGTGCAGTGCACCCACGG
CATCCGCCCCGTGGTGAGCACCCAGCTGCTGCTGAACGGCAGCCTGGCCG
AGGAGGAGGTGGTGATCCGCAGCGAGAACTTCACCGACAACGCCAAGACC
ATCATCGTGCAGCTGAACGAGAGCGTGGAGATCAACTGCACGCGTCCCAA
CAACAACACCCGCAAGAGCATCCCCATCGGCCCTGGCCGCGCCCTGTACG
CCACCGGCAAGATCATCGGCGACATCCGCCAGGCCCACTGCAACCTGTCG
CGAGCCAAGTGGAACAACACCCTGAAGCAGATCGTGACCAAGCTGCGCGA
GCAGTTCGGCAACAACAAGACCACCATCGTGTTCAACCAGAGCAGCGGCG
GCGACCCCGAGATCGTGATGCACAGCTTCAACTGCGGCGGCGAATTCTTC
TACTGCAACAGCACCCAGCTGTTCAACAGCACCTGGCACTTCAACGGCAC
CTGGGGCAACAACAACACCGAGCGCAGCAACAACGCCGCCGACGACAACG
ACACCATCACCCTGCCCTGCCGCATCAAGCAGATCATCAACATGTGGCAG
GAGGTGGGCAAGGCCATGTACGCCCCCCCCATCAGCGGCCAGATCCGCTG
CAGCAGCAACATCACCGGCCTGCTGCTGACTCGAGACGGCGGCAACAACG
AGAACACCAACAACACCGACACCGAGATCTTCCGCCCCGGGGGCGGCGAC
ATGCGCGACAACTGGCGCAGCGAGCTGTACAAGTACAAGGTGGTGAAGAT
CGAGCCCCTGGGCGTAGCACCCACCAAGGCAAAGAGAAGAGTGGTGCAGA
GAGAAAAAAGCGCAGTGGGAATGCTAGGAGCTATGTTCCTTGGGTTCTTG
GGAGCAGCAGGAAGCACTATGGGCGCAGCGTCAATGACGCTGACCGTACA
GGCCAGACAATTATTGTCTGGTATAGTGCAGCAGCAGAACAATCTGCTGA

GGGCTATTGAGGCGCAACAGCATCTGTTGCAACTCACAGTCTGGGGCATC
AAGCAGCTCCAGGCAAGAGTCCTGGCTGTGGAAAGATACCTAAAGGATCA
GCAGCTCCTGGGGATTTGGGGTTGCTCTGGAAAACTCATCTGCACCACTG
CTGTGCCTTGGAATGCTAGCTGGAGCAACAAGAGCCTGGACAAGATCTGG
AACAACATGACCTGGATGGAGTGGGAGCGCGAGATCGACAACTACACCGG
CCTGATCTACACCCTGATCGAGGAGAGCCAGAACCAGCAGGAGAAGAACG
AGCAGGAGCTGCTGGAGCTGGACAAGTGGGCCAGCCTGTGGAACTGGTTC
GATATCACCAACTGGCTGTGGTACATCAAGATCTTCATCATGATCGTGGG
CGGCCTGGTGGGCCTGCGCATCGTGTTCGCCGTGCTGAGCATCGTGAACC
GCGTGCGCCAGGGCTACAGCCCCCTGAGCTTCCAGACCCACCTGCCAGCC
CCGAGGGGACCCGACAGGCCCGAAGGAATCGAAGAAGAAGGTGGAGAGAG
AGACAGAGACAGATCCGGTCGATTAGTGAATGGATTCTTAGCACTTATCT
GGGACGACCTGCGGAGCCTGTGCCTCTTCAGCTACCACCGCTTGAGCGAC
TTACTCTTGATTGTAGCGAGGATTGTGGAACTTCTGGGACGCAGGGGGTG
GGAGGCCCTCAAATATTGGTGGAATCTCCTGCAGTACTGGAGTCAGGAAC
TAAAGAATAGCGCCGTGAGCCTGCTGAACGCCACCGCCATCGCCGTGGCC
GAGGGCACCGACCGCGTGATCGAGGTGGTGCAGCGCGCCTGCCGCGCCAT
CCTGCACATCCCCCGCCGCATCCGCCAGGGCCTGGAGCGCGCCCTGCTGT
GA

Table 6 (SEQ ID NO:6)

ATGAGAGTGATGGGGATACTGAGGAATTGTCAACAATGGTGGATATGGGG
CATCCTAGGCTTTTGGATGCTAATGATTTGTGACGTGATGGGCAACCTGT
GGGTGACCGTGTACTACGGCGTGCCCGTGTGGAAGGAGGCCAAGACCACC
CTGTTCTGCGCCAGCGACGCCAAGGCCTACGAGCGGGAGGTGCACAACGT
GTGGGCCACCCACGCCTGCGTGCCCACCGACCCCAACCCCCAGGAGATGG
TGCTGGAGAACGTGACCGAGAACTTCAACATGTGGAAGAACGACATGGTG
GACCAGATGCACGAGGACATCATCAGCCTGTGGGACCAGAGCCTGAAGCC
CTGCGTGAAGCTGACCCCCCTGTGCGTGACCCTGAACTGCACCAACGTGA
CCAACACCAACAACAACAACAACACCAGCATGGGCGGCGAGATCAAGAAC
TGCAGCTTCAACATCACCACCGAGCTGCGGGACAAGAAGCAGAAGGTGTA
CGCCCTGTTCTACCGGCTGGACATCGTGCCCCTGAACGAGAACAGCAACA
GCAACAGCAGCGAGTACCGGCTGATCAACTGCAACACCAGCGCCATCACC
CAGGCCTGCCCCAAGGTGAGCTTCGACCCCATCCCCATCCACTACTGCGC
CCCCGCCGGCTACGCCATCCTGAAGTGCAACAACAAGACCTTCAACGGCA
CCGGCCCCTGCAACAACGTGAGCACCGTGCAGTGCACCCACGGCATCAAG
CCCGTGGTGAGCACCCAGCTGCTGCTGAACGGCAGCCTGGCCGAGGAGGA
GATCATCATCCGGAGCGAGAACCTGACCAACAACGCCAAGACCATCATCG
TGCACCTGAACGAGAGCGTGGAGATCGTGTGCACCCGGCCCAACAACAAC
ACCCGGAAGAGCATCCGGATCGGCCCCGGCCAGACCTTCTACGCCACCGG
CGACATCATCGGCGACATCCGGCAGGCCCACTGCAACATCAGCGAGAAGG
AGTGGAACAAGACCCTGCAGCGGGTGGGCAAGAAGCTGAAGGAGCACTTC
CCCAACAAGACCATCAAGTTCGAGCCCAGCAGCGGCGGCGACCTGGAGAT
CACCACCCACAGCTTCAACTGCCGGGGCGAGTTCTTCTACTGCAACACCA
GCAAGCTGTTCAACAGCACCTACAACAGCACCAACAACGGCACCACCAGC
AACAGCACCATCACCCTGCCCTGCCGGATCAAGCAGATCATCAACATGTG
GCAGGGCGTGGGCCGGGCCATGTACGCCCCCCCCATCGCCGGCAACATCA
CCTGCAAGAGCAACATCACCGGCCTGCTGCTGACCCGGGACGGCGGCAAC
ACCAACAACACCACCGAGACCTTCCGGCCCGGCGGCGGCGACATGCGGGA
CAACTGGCGGAGCGAGCTGTACAAGTACAAGGTGGTGGAGATCAAGCCCC
TGGGCGTAGCACCCACTGAGGCAAAAAGGAGAGTGGTGGAGAGAGAAAAA
AGAGCAGTGGGAATAGGAGCTGTGTTCCTTGGGTTCTTGGGAGCAGCAGG
AAGCACTATGGGCGCGGCGTCAATAACGCTGACGGTACAGGCCAGACAAT
TATTGTCTGGTATAGTGCAACAGCAAAGCAATTTGCTGAGGGCTATAGAG
GCGCAACAGCATATGTTGCAACTCACGGTCTGGGGCATTAAGCAGCTCCA
GACAAGAGTCCTGGCTATAGAAAGATACCTAAAGGATCAGCAGCTCCTGG
GCATTTGGGGCTGCTCTGGAAAACTCATCTGCACCACTGCTGTGCCTTGG
AACTCTAGCTGGAGCAACAAGAGCCAGGACGACATCTGGGACAACATGAC
CTGGATGCAGTGGGACCGGGAGATCAGCAACTACACCGACACCATCTACC

GGCTGCTGGAGGACAGCCAGAACCAGCAGGAGAAGAACGAGAAGGACCTG
CTGGCCCTGGACAGCTGGAAGAACCTGTGGAACTGGTTCGACATCACCAA
CTGGCTGTGGTACATCAAGATCTTCATCATGATCGTGGGCGGCCTGATCG
GCCTGCGGATCATCTTCGCCGTGCTGAGCATCGTGAACCGGGTGCGGCAG
GGCTACAGCCCCCTGAGCTTCCAGACCCTTACCCCAAACCCGAGGGGACC
CGACAGGCTCGGAGGAATCGAAGAAGAAGGTGGAGAGCAAGACAGAGACA
GATCCATTCGATTAGTGAGCGGATTCTTAGCACTGGCCTGGGACGACCTG
CGGAGCCTGTGCCTCTTCAGCTACCACCGATTGAGAGACTTCATATTGAT
TGCAGCCAGAGGGTGGGAACTTCTGGGACGCAGCAGTCTCAGGGGACTGC
AGAGGGGGTGGGAAGCCCTTAAGTATCTGGGAAGTCTTGTGCAGTATTGG
GGTCTGGAGCTAAAAAAGAGTGCTATTAGCCTGCTGGACACCATCGCCAT
CGCCGTGGCCGAGGGCACCGACCGGATCATCGAGCTGGTGCAGCGGATCT
GCCGGGCCATCCGGAACATCCCCCGGCGGATCCGGCAGGGCTTCGAGGCC
GCCCTGCAGTGA

Unless otherwise stated, all enzymes are used in accordance with the manufacturer's instructions.

[0183] In a typical embodiment, a nucleic acid encoding the ancestral or COT viral sequence is synthesized by joining long oligonucleotides. By synthesizing a nucleic acid *de novo*, desired features are easily incorporated into the gene. Such features include, but are not limited to, the incorporation of convenient restriction sites to enable further manipulation of the nucleic acid sequence, optimization of the codon frequencies (*e.g.*, human codon frequencies) to greatly enhance *in vivo* expression levels, which can favor the immunogenicity of the polypeptide sequence, and the like. Long oligonucleotides can be synthesized with a very low error rate using the solid-phase method. Long oligonucleotides designed with a 20-25 nucleotide complementary sequence at both 5' and 3' ends can be joined using DNA polymerase, DNA ligase, and the like. If necessary, the sequence of the synthesized nucleic acid can be verified by DNA sequence analysis.

[0184] Oligonucleotides that are not commercially available can be chemically synthesized. Suitable methods include, for example, the solid phase phosphoramidite triester method first described by Beaucage and Caruthers (*Tetrahedron Letts* 22(20):1859-62 (1981)), and the use of an automated synthesizer (*see, e.g.*, Needham Van Devanter *et al.*, *Nucleic Acids Res.* 12:6159-68 (1984)). Purification of oligonucleotides is, for example, by native acrylamide gel electrophoresis or by anion-exchange HPLC, as described in Pearson and Reanier (*J. Chrom.* 255:137-49 (1983)).

[0185] The sequence of the nucleic acids can be verified, for example, using the chemical degradation method of Maxam *et al.* (*Methods in Enzymology* 65:499-560 (1980)), or the chain termination method for sequencing double stranded templates (*see, e.g.*, Wallace *et al.*, *Gene* 16:21-26 (1981)). Southern blot hybridization techniques can be carried out according to Southern *et al.* (*J. Mol. Biol.* 98:503 (1975)), Sambrook *et al.* (*supra*), or Ausubel *et al.* (*supra*).

[0186] *Expression of Ancestral or COT Viral Sequences*

[0187] The nucleic acids encoding ancestral or COT viral sequences can be inserted into an appropriate expression vector (*i.e.*, a vector which contains the necessary elements for the transcription and translation of the inserted polypeptide-coding sequence). A variety of host-vector systems can be utilized to express the polypeptide-coding sequence(s). These include, for example, mammalian cell systems infected with virus (*e.g.*, vaccinia virus, adenovirus,

sindbis virus, Venezuelan equine encephalitis (VEE) virus, and the like), insect cell systems infected with virus (*e.g.*, baculovirus), microorganisms such as yeast containing yeast vectors, or bacteria transformed with bacteriophage DNA, plasmid DNA, or cosmid DNA. The expression elements of vectors vary in their strengths and specificities. Depending on the host-vector system utilized, any one of a number of suitable transcription and translation elements can be used. In specific embodiments, the ancestral or COT viral sequence is expressed in human cells, other mammalian cells, yeast or bacteria. In yet another embodiment, a fragment of an ancestral or COT viral sequence comprising an immunologically active region of the sequence is expressed.

[0188] Any suitable method can be used for insertion of nucleic acids encoding ancestral or COT viral sequences into an expression vector. Suitable expression vectors typically include appropriate transcriptional and translational control signals. Suitable methods include *in vitro* recombinant DNA and synthetic techniques and *in vivo* recombination techniques (genetic recombination). Expression of nucleic acid sequences can be regulated by a second nucleic acid sequence so that the encoded nucleic acid is expressed in a host transformed with the recombinant DNA molecule. For example, expression of an ancestral or COT viral sequence can be controlled by any suitable promoter/enhancer element known in the art. Suitable promoters include, for example, the SV40 early promoter region (Benoist and Chambon, *Nature* 290:304-10 (1981)), the promoter contained in the 3' long terminal repeat of Rous sarcoma virus (Yamamoto *et al.*, *Cell* 22:787-97 (1980)), the herpes thymidine kinase promoter (Wagner *et al.*, *Proc. Natl. Acad. Sci. USA* 78:1441-45 (1981)), the Cytomegalovirus promoter, the translational elongation factor EF-1α promoter, the regulatory sequences of the metallothionein gene (Brinster *et al.*, *Nature* 296:39-42 (1982)), prokaryotic promoters such as, for example, the β-lactamase promoter (Villa-Komaroff *et al.*, *Proc. Natl. Acad. Sci. USA* 75:3727-31 (1978)) or the *tac* promoter (deBoer *et al.*, *Proc. Natl. Acad. Sci. USA* 80:21-25 (1983)), plant expression vectors including the cauliflower mosaic virus 35S RNA promoter (Gardner *et al.*, *Nucl. Acids Res.* 9:2871-88 (1981)), and the promoter of the photosynthetic enzyme ribulose biphosphate carboxylase (Herrera-Estrella *et al.*, *Nature* 310:115-20 (1984)), promoter elements from yeast or other fungi such as the *GAL7* and *GAL4* promoters, the *ADH* (alcohol dehydrogenase) promoter, the *PGK* (phosphoglycerol kinase) promoter, the alkaline phosphatase promoter, and the like.

[0189] Other exemplary mammalian promoters include, for example, the following animal transcriptional control regions, which exhibit tissue specificity: the elastase I gene control region which is active in pancreatic acinar cells (Swift *et al.*, *Cell* 38:639-46 (1984); Ornitz *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* 50:399-409 (1986); MacDonald, *Hepatology* 7(1 Suppl.):42S-51S (1987); the insulin gene control region which is active in pancreatic beta cells (Hanahan, *Nature* 315:115-22 (1985)), the immunoglobulin gene control region which is active in lymphoid cells (Grosschedl *et al.*, *Cell* 38:647-58 (1984); Adams *et al.*, *Nature* 318:533-38 (1985); Alexander *et al.*, *Mol. Cell. Biol.* 7:1436-44 (1987)), the mouse mammary tumor virus control region which is active in testicular, breast, lymphoid and mast cells (Leder *et al.*, *Cell* 45:485-95 (1986)), the albumin gene control region which is active in liver (Pinkert *et al.*, *Genes Dev.* 1:268-76 (1987)), the alpha-fetoprotein gene control region which is active in liver (Krumlauf *et al.*, *Mol. Cell. Biol.* 5:1639-48 (1985); Hammer *et al.*, *Science* 235:53-58 (1987); the alpha 1-antitrypsin gene control region which is active in the liver (Kelsey *et al.*, *Genes and Devel.* 1:161-71 (1987)); the beta-globin gene control region which is active in myeloid cells (Magram *et al.*, *Nature* 315:338-40 (1985); Kollias *et al.*, *Cell* 46:89-94 (1986); the myelin basic protein gene control region which is active in oligodendrocyte cells in the brain (Readhead *et al.*, *Cell* 48:703-12 (1987)); the myosin light chain-2 gene control region which is active in skeletal muscle (Shani, *Nature* 314:283-86 (1985)); and the gonadotropic releasing hormone gene control region which is active in the hypothalamus (Mason *et al.*, *Science* 234:1372-78 (1986)).

[0190] In a specific embodiment, a vector is used that comprises a promoter operably linked to the ancestral or COT viral sequence encoding nucleic acid, one or more origins of replication, and, optionally, one or more selectable markers (*e.g.*, an antibiotic resistance gene). Suitable selectable markers include, for example, those conferring resistance to ampicillin, tetracycline, neomycin, G418, and the like. An expression construct can be made, for example, by subcloning a nucleic acid encoding an ancestral or COT viral sequence into a restriction site of the pRSECT expression vector. Such a construct allows for the expression of the ancestral or COT viral sequence under the control of the T7 promoter with a histidine amino terminal flag sequence for affinity purification of the expressed polypeptide.

[0191] In an exemplary embodiment, a high efficiency expression system can be used which employs a high-efficiency DNA transfer vector (the pJW4304 SV40/EBV vector) with

a very high efficiency RNA/protein expression component (*e.g.*, from the Semliki Forest Virus) to achieve maximal protein expression, as further discussed *infra*. pJW4304 SV40/EBV was prepared from pJW4303, which is described by Robinson *et al.* (*Ann. New York Acad. Sci.* 27:209-11 (1995)) and Yasutomi *et al.* (*J. Virol.* 70:678-81 (1996)).

[0192] Expression vector/host systems expressing an ancestral or COT viral sequences can be identified by general approaches well known to the skilled artisan, including: (a) nucleic acid hybridization, (b) the presence or absence of "marker" gene function, (c) expression of inserted sequences; or (d) screening transformed cells by standard recombinant DNA methods. In the first approach, the presence of an ancestral or COT viral sequence nucleic acid inserted in host cells can be detected by nucleic acid hybridization using probes comprising sequences that are homologous to an inserted nucleic acid. In the second approach, the expression vector/host system can be identified and selected based upon the presence or absence of certain "marker" gene functions (*e.g.*, thymidine kinase activity, resistance to antibiotics, transformation phenotype, occlusion body formation in baculovirus, and the like) caused by the insertion of a vector containing the desired nucleic acids. For example, if the nucleic acid is inserted within the marker gene sequence of the vector, recombinants containing the ancestral or COT viral sequence can be identified by the absence of the marker gene function.

[0193] In the third approach, expression vector/host systems can be identified by assaying for the ancestral or COT viral sequence polypeptide expressed by the recombinant host organism. Such assays can be based, for example, on the physical or functional properties of the ancestral or COT viral sequence polypeptide in *in vitro* assay systems (*e.g.*, binding by antibody). In the fourth approach, expression vector/host cells can be identified by screening transformed host cells by known recombinant DNA methods.

[0194] Once a suitable expression vector host system and growth conditions are established, methods that are known in the art can be used to propagate it. In addition, host cells can be chosen that modulate the expression of the inserted nucleic acid sequences, or that modify or process the gene product in the specific fashion desired. Expression from certain promoters can be elevated in the presence of certain inducers; thus, expression of the ancestral or COT viral sequence can be controlled. Furthermore, different host cells having characteristic and specific mechanisms for the translational and post-translational processing

and modification (*e.g.*, glycosylation or phosphorylation) of polypeptides can be used. Appropriate cell lines or host systems can be chosen to ensure the desired modification and processing of the expressed polypeptide. For example, expression in a bacterial system can be used to produce an unglycosylated polypeptide.

[0195] *Antibodies to Ancestor or COT Proteins, Fragments, Derivatives and Analogs:*

[0196] Ancestor or COT proteins (including fragments, derivatives, and analogs thereof), can be used as an immunogen to generate antibodies which immunospecifically bind such ancestor or COT proteins and to circulating variants. Such antibodies include but are not limited to polyclonal antibodies, monoclonal antibodies, chimeric antibodies, single chain antibodies, antigen binding antibody fragments (*e.g.*, Fab, Fab', $F(ab')_2$, Fv, or hypervariable regions), and an Fab expression library. In some embodiments, polyclonal and/or monoclonal antibodies to an ancestor or COT protein are produced. In other embodiments, antibodies to a domain of an ancestor or COT protein are produced. In yet other embodiments, fragments of an ancestor or COT protein that are identified as immunogenic (*e.g.*, hydrophilic) are used as immunogens for antibody production.

[0197] Various procedures known in the art can be used for the production of polyclonal antibodies. For the production of such antibodies, various host animals (including, but not limited to, rabbits, mice, rats, sheep, goats, camels, and the like) can be immunized by injection with the ancestor or COT protein, fragment, derivative or analog. Various adjuvants can be used to increase the immunological response, depending on the host species including, but not limited to, Freund's adjuvant (complete and incomplete), mineral gels such as aluminum hydroxide, surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanins, dinitrophenol, and potentially useful human adjuvants such as BCG (bacille Calmette-Guerin) and *Corynebacterium parvum.*

[0198] For preparation of monoclonal antibodies directed toward an ancestor or COT protein, fragment, derivative, or analog thereof, any technique that provides for the production of antibody molecules by continuous cell lines in culture can be used. Such techniques include, for example, the hybridoma technique originally developed by Kohler

and Milstein (*see, e.g., Nature* 256:495-97 (1975)), the trioma technique (*see, e.g.*, Hagiwara and Yuasa, *Hum. Antibodies Hybridomas.* 4:15-19 (1993); Hering *et al., Biomed. Biochim. Acta* 47:211-16 (1988)), the human B-cell hybridoma technique (*see, e.g.*, Kozbor *et al., Immunology Today* 4:72 (1983)), and the EBV-hybridoma technique to produce human monoclonal antibodies (*see, e.g.*, Cole *et al.*, In: *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, Inc., pp. 77-96 (1985)). Human antibodies can be used and can be obtained by using human hybridomas (*see, e.g.*, Cote *et al., Proc. Natl. Acad. Sci. USA* 80:2026-30 (1983)) or by transforming human B cells with EBV virus *in vitro* (*see, e.g.*, Cole *et al., supra*).

[0199] Further to the invention, "chimeric" or "humanized" antibodies (*see, e.g.*, Morrison *et al., Proc. Natl. Acad. Sci. USA* 81:6851-55 (1984); Neuberger *et al., Nature* 312:604-08 (1984); Takeda *et al., Nature* 314:452-54 (1985)) can be prepared. Such chimeric antibodies are typically prepared by splicing the non-human genes for an antibody molecule specific for ancestor or COT protein together with genes from a human antibody molecule of appropriate biological activity. It can be desirable to transfer the antigen binding regions (*e.g.*, Fab', $F(ab')_2$, Fab , Fv, or hypervariable regions) of non-human antibodies into the framework of a human antibody by recombinant DNA techniques to produce a substantially human molecule. Methods for producing such "chimeric" molecules are generally well known and described in, for example, U.S. Patent Nos. 4,816,567; 4,816,397; 5,693,762; and 5,712,120; International Patent Publications WO 87/02671 and WO 90/00616; and European Patent Publication EP 239 400 (the disclosures of which are incorporated by reference herein). Alternatively, a human monoclonal antibody or portions thereof can be identified by first screening a human B-cell cDNA library for DNA molecules that encode antibodies that specifically bind to an ancestor or COT protein according to the method generally set forth by Huse *et al.* (*Science* 246:1275-81 (1989)). The DNA molecule can then be cloned and amplified to obtain sequences that encode the antibody (or binding domain) of the desired specificity. Phage display technology offers another technique for selecting antibodies that bind to ancestor or COT proteins, fragments, derivatives or analogs thereof. (*See, e.g.*, International Patent Publications WO 91/17271 and WO 92/01047; Huse *et al., supra.*)

[0200] According to another aspect of the invention, techniques described for the production of single chain antibodies (*see, e.g.*, U.S. Patents Nos. 4,946,778 and 5,969,108)

can be adapted to produce single chain antibodies. An additional aspect of the invention utilizes the techniques described for the construction of a Fab expression library (*see, e.g.*, Huse *et al.*, *supra*) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity for ancestor or COT proteins, fragments, derivatives, or analogs thereof.

[0201] Antibody that contains the idiotype of the molecule can be generated by known techniques. For example, such fragments include but are not limited to, the $F(ab')_2$ fragment which can be produced by pepsin digestion of the antibody molecule, the Fab' fragments which can be generated by reducing the disulfide bridges of the $F(ab')_2$ fragment, the Fab fragments which can be generated by treating the antibody molecule with papain and a reducing agent, and Fv fragments. Recombinant Fv fragments can also be produced in eukaryotic cells using, for example, the methods described in U.S. Patent No. 5,965,405.

[0202] In the production of antibodies, screening for the desired antibody can be accomplished by techniques known in the art (*e.g.*, ELISA (enzyme-linked immunosorbent assay)). In one example, antibodies that recognize a specific domain of an ancestor or COT protein can be used to assay generated hybridomas for a product which binds to polypeptide containing that domain. Antibodies specific to a domain of an ancestor or COT protein are also provided.

[0203] Antibodies against ancestor or COT proteins (including fragments, derivatives and analogs) can be used for passive antibody treatment, according to methods known in the art. Antibodies can be introduced into an individual to prevent or treat viral infection. Typically, such antibody therapy is practiced as an adjuvant to the vaccination protocols. The antibodies can be produced as described *supra* and can be polyclonal or monoclonal antibodies and administered intravenously, enterally (*e.g.*, as an enteric coated tablet form), by aerosol, orally, transdermally, transmucosally, intrapleurally, intrathecally, or by other suitable routes.

[0204] *Immunogenic Compositions and Vaccines*

[0205] The present invention also provides immunogenic compositions, such as vaccines. An example of the development of a vaccine ("digital vaccine") using the sequences of the invention is illustrated in Figure 4. The present invention also provides a new way to

produce vaccines, using HIV ancestral or COT viral sequences (*e.g.*, HIV *env* or *gag* genes or polypeptides). Such ancestral or COT viral sequences typically correspond to the structure of a real biological entity - the founding virus (*i.e.*, "the viral Eve").

[0206] *Formulations*

[0207] Immunogenic compositions and vaccines that contain an immunogenically effective amount of one or more ancestor or COT viral protein sequences, or fragments, derivatives, or analogs thereof, are provided. Immunogenic epitopes in an ancestral or COT protein sequence can be identified according to methods known in the art, and proteins, fragments, derivatives, or analogs containing those epitopes can be delivered by various means, in a vaccine composition. Suitable compositions can include, for example, lipopeptides (*e.g.*, Vitiello *et al.*, *J. Clin. Invest.* 95:341 (1995)), peptide compositions encapsulated in poly(DL-lactide-co-glycolide) ("PLG") microspheres (*see, e.g.*, Eldridge *et al.*, *Molec. Immunol.* 28:287-94 (1991); Alonso *et al.*, *Vaccine* 12:299-306 (1994); Jones *et al.*, *Vaccine* 13:675-81 (1995)), peptide compositions contained in immune stimulating complexes (ISCOMS) (*see, e.g.*, Takahashi *et al.*, *Nature* 344:873-75 (1990); Hu *et al.*, *Clin. Exp. Immunol.* 113:235-43 (1998)), multiple antigen peptide systems (MAPs) (*see, e.g.*, Tam, *Proc. Natl. Acad. Sci. U.S.A.* 85:5409-13 (1988); Tam, *J. Immunol. Methods* 196:17-32 (1996)), viral delivery vectors (*see, e.g.*, Perkus *et al.*, In: *Concepts in vaccine development*, Kaufmann (ed.), p. 379 (1996)), particles of viral or synthetic origin (*see, e.g.*, Kofler *et al.*, *J. Immunol. Methods.* 192:25-35 (1996); Eldridge *et al.*, *Sem. Hematol.* 30:16 (1993); Falo *et al.*, *Nature Med.* 7:649 (1995)), adjuvants (*see, e.g.*, Warren *et al.*, *Annu. Rev. Immunol.* 4:369 (1986); Gupta *et al.*, *Vaccine* 11:293 (1993)), liposomes (*see, e.g.*, Reddy *et al.*, *J. Immunol.* 148:1585 (1992); Rock, *Immunol. Today* 17:131 (1996)), or naked or particle absorbed cDNA (*see, e.g.*, Shiver *et al.*, In: *Concepts in vaccine development*, Kaufmann (ed.), p. 423 (1996)). Toxin-targeted delivery technologies, also known as receptor-mediated targeting, such as those of Avant Immunotherapeutics, Inc. (Needham, Massachusetts) can also be used.

[0208] Furthermore, useful carriers that can be used with immunogenic compositions and vaccines of the invention are well known in the art, and include, for example, thyroglobulin, albumins such as human serum albumin, tetanus toxoid, polyamino acids such as poly L-lysine, poly L-glutamic acid, influenza, hepatitis B virus core protein, and the like. The compositions and vaccines can contain a physiologically tolerable (*i.e.*, acceptable) diluent

such as water, or saline, typically phosphate buffered saline. The compositions and vaccines also typically include an adjuvant. Adjuvants such as incomplete Freund's adjuvant, aluminum phosphate, aluminum hydroxide, or alum are examples of materials well known in the art. Additionally, as disclosed herein, CTL responses can be primed by conjugating ancestor or COT proteins (or fragments, derivative or analogs thereof) to lipids, such as tripalmitoyl-S-glycerylcysteinyl-seryl- serine (P_3CSS).

[0209] As disclosed in greater detail herein, upon immunization with a composition or vaccine containing an ancestor or COT viral sequence protein composition in accordance with the invention, via injection, aerosol, oral, transdermal, transmucosal, intrapleural, intrathecal, or other suitable routes, the immune system of the host responds to the composition or vaccine by producing large amounts of CTL's, HTL's and/or antibodies specific for the desired antigen. Consequently, the host typically becomes at least partially immune to later infection, or at least partially resistant to developing an ongoing chronic infection, or derives at least some therapeutic benefit.

[0210] For therapeutic or prophylactic immunization, ancestor or COT proteins (including fragments, derivatives and analogs) can also be expressed by viral or bacterial vectors. Examples of expression vectors include attenuated viral hosts, such as vaccinia or fowlpox. In one embodiment, this approach involves the use of vaccinia virus, for example, as a vector to express nucleotide sequences that encode the polypeptide. Upon introduction into an acutely or chronically infected host, or into a non-infected host, the recombinant vaccinia virus expresses the immunogenic protein, and thereby elicits a host CTL, HTL and/or antibody response. Vaccinia vectors and methods useful in immunization protocols are described in, for example, U.S. Patent No. 4,722,848, the disclosure of which is incorporated by reference herein. A wide variety of other vectors useful for therapeutic administration or immunization of the peptides of the invention, for example, adeno and adeno-associated virus vectors, retroviral vectors, *Salmonella typhimurium* vectors, detoxified anthrax toxin vectors, Alphavirus, and the like, can also be used, as will be apparent to those skilled in the art from the description herein. Alphavirus vectors that can be used include, for example, Sindbis and Venezuelan equine encephalitis (VEE) virus. (*See, e.g.*, Coppola *et al.*, *J. Gen. Virol.* 76:635-41 (1995); Caley *et al.*, *Vaccine* 17:3124-35 (1999); Loktev *et al.*, *J. Biotechnol.* 44:129-37 (1996).)

[0211] Polynucleotides (*e.g.*, DNA or RNA) encoding one or more ancestral or COT proteins (including fragments, derivative or analogs) can also be administered to a patient. This approach is described in, for example, Wolff *et al.*, (*Science* 247:1465 (1990)), in U.S. Patent Nos. 5,580,859; 5,589,466; 5,804,566; 5,739,118; 5,736,524; 5,679,647; and WO 98/04720; and in more detail below. Examples of DNA-based delivery technologies include "naked DNA", facilitated (bupivicaine, polymer, or peptide-mediated) delivery, cationic lipid complexes, particle-mediated ("gene gun"), or pressure-mediated delivery (*see*, *e.g.*, U.S. Patent No. 5,922,687).

[0212] The direct injection of naked plasmid DNA encoding a protein antigen as a means of vaccination is, among several HIV delivery and expression systems that have been developed in the last decade, one that has attracted much attention. In mouse models, as well as in large animal models, both humoral and cellular immune responses are readily induced, resulting in protective immunity against challenge infections in some instances. A Semliki Forest Virus (SFV) replicon can also be used, for example, in the context of naked DNA immunization. SFV belongs to the Alphavirus family wherein the genome consists of a single stranded RNA of positive polarity encoding its own replicase. By replacing the SFV structural genes with the gene of interest, expression levels as high as 25% of the total cell protein are obtained. Another advantage of this alphavirus over plasmid vectors is its non-persistence: the antigen of interest is expressed at high levels but for a short period (typically <72 hours). In contrast, plasmid vectors generally induce synthesis of the antigen of interest over extended time periods, risking chromosomal integration of foreign DNA and cell transformation. Furthermore, antigen persistence or repeated inoculations of small amounts of antigen has been shown experimentally to induce tolerance. Prolonged antigen synthesis, therefore, can theoretically result in unresponsiveness rather than immunity.

[0213] Ancestor or COT proteins, fragments, derivative, and analogs can also be introduced into a subject *in vivo* or *ex vivo*. For example, ancestral or COT viral sequences can be transferred into defined cell populations. Suitable methods for gene transfer include, for example:

[0214] 1) Direct gene transfer. (*See*, *e.g.*, Wolff *et al.*, *Science* 247:1465-68 (1990)).

[0215] 2) Liposome-mediated DNA transfer. (*See*, *e.g.*, Caplen *et al.*, *Nature Med.* 3:39-46

(1995); Crystal, *Nature Med.* 1:15-17 (1995); Gao and Huang, *Biochem. Biophys. Res. Comm.* 179:280-85 (1991).)

[0216] 3) Retrovirus-mediated DNA transfer. (*See, e.g.*, Kay *et al.*, *Science* 262:117-19 (1993); Anderson, *Science* 256:808-13 (1992).) Retroviruses from which the retroviral plasmid vectors can be derived include lentiviruses. They further include, but are not limited to, Moloney Murine Leukemia Virus, spleen necrosis virus, retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, gibbon ape leukemia virus, human immunodeficiency virus, Myeloproliferative Sarcoma Virus, and mammary tumor virus. In one embodiment, the retroviral plasmid vector is derived from Moloney Murine Leukemia Virus. Examples illustrating the use of retroviral vectors in gene therapy further include the following: Clowes *et al.* (*J. Clin. Invest.* 93:644-51 (1994)); Kiem *et al.* (*Blood* 83:1467-73 (1994)); Salmons and Gunzberg (*Human Gene Therapy* 4:129-41 (1993)); and Grossman and Wilson (*Curr. Opin. in Genetics and Devel.* 3:110-14 (1993)).

[0217] 4) DNA Virus-mediated DNA transfer. Such DNA viruses include adenoviruses (*e.g.*, Ad-2 or Ad-5 based vectors), herpes viruses (typically herpes simplex virus based vectors), and parvoviruses (*e.g.*, "defective" or non-autonomous parvovirus based vectors, or adeno-associated virus based vectors, such as AAV-2 based vectors). (*See, e.g.*, Ali *et al.*, *Gene Therapy* 1:367-84 (1994); U.S. Patent Nos. 4,797,368 and 5,139,941, the disclosures of which are incorporated herein by reference.) Adenoviruses have the advantage that they have a broad host range, can infect quiescent or terminally differentiated cells, such as neurons or hepatocytes, and appear essentially non-oncogenic. Adenoviruses do not appear to integrate into the host genome. Because they exist extrachromosomally, the risk of insertional mutagenesis is greatly reduced. Adeno-associated viruses exhibit similar advantages as adenoviral-based vectors. However, AAVs exhibit site-specific integration on human chromosome 19.

[0218] Kozarsky and Wilson (*Current Opinion in Genetics and Development* 3:499-503 (1993)) present a review of adenovirus-based gene therapy. Bout *et al.* (*Human Gene Therapy* 5:3-10 (1994)) demonstrated the use of adenovirus vectors to transfer genes to the respiratory epithelia of rhesus monkeys. Herman *et al.* (*Human Gene Therapy* 10:1239-49 (1999)) describe the intraprostatic injection of a replication-deficient adenovirus containing the herpes simplex thymidine kinase gene into human prostate, followed by intravenous

administration of the prodrug ganciclovir in a phase I clinical trial. Other instances of the use of adenoviruses in gene therapy can be found in Rosenfeld *et al.* (*Science* 252:431-34 (1991)); Rosenfeld *et al.* (*Cell* 68:143-55 (1992)); Mastrangeli *et al.* (*J. Clin. Invest.* 91:225-34 (1993)); Thompson (*Oncol. Res.* 11:1-8 (1999)).

[0219] The choice of a particular vector system for transferring the ancestral or COT viral sequence of interest will depend on a variety of factors. One important factor is the nature of the target cell population. Although retroviral vectors have been extensively studied and used in a number of gene therapy applications, these vectors are generally unsuited for infecting non-dividing cells. In addition, retroviruses have the potential for oncogenicity. However, recent developments in the field of lentiviral vectors may circumvent some of these limitations. (*See* Naldini *et al.*, *Science* 272:263-67 (1996).)

[0220] The skilled artisan will appreciate that any suitable expression vector containing nucleic acid encoding an ancestor or COT protein, or fragment, derivative or analog thereof can be used in accordance with the present invention. Techniques for constructing such a vector are known. (*See, e.g.*, Anderson, *Nature* 392:25-30 (1998); Verma, *Nature* 389:239-42 (1998).) Introduction of the vector to the target site can be accomplished using known techniques.

[0221] In another one embodiment, a novel expression system employing a high-efficiency DNA transfer vector (the pJW4304 SV40/EBV vector (pJW4304 SV40/EBV was prepared from pJW4303, which is described by Robinson *et al.*, *Ann. New York Acad. Sci.* 27:209-11 (1995) and Yasutomi *et al.*, *J. Virol.* 70:678-81 (1996)) with a very high efficiency RNA/protein expression system (the Semliki Forest Virus) is used to achieve maximal protein expression in vaccinated hosts with a safe and inexpensive vaccine. SFV cDNA is placed, for example, under the control of a cytomegalovirus (CMV) promoter (*see* Figure 7). Unlike conventional DNA vectors, the CMV promoter does not directly drive the expression of the antigen encoding nucleic acids. Instead, it directs the synthesis of recombinant SFV replicon RNA transcript. Translation of this RNA molecule produces the SFV replicase complex, which catalyzes cytoplasmic self-amplification of the recombinant RNA, and eventual high-level production of the actual antigen-encoding mRNA. Following vector delivery, the transfected host cell dies within a few days. In the context of the present invention, *env* and/or *gag* genes are typically cloned into this vector. *In vitro* experiments

using Northern blot, Western blot, SDS-PAGE, immunoprecipitation assay, and CD4 binding assays can be performed, as described *infra*, to determine the efficiency of this system by assessing protein expression level, protein characteristics, duration of expression, and cytopathic effects of the vector.

[0222] In some embodiments, ancestor or COT protein (or a fragment, derivative or analog thereof) is administered to a subject in need thereof. The dosage for an initial therapeutic immunization generally occurs in a unit dosage range where the lower value is about 1, 5, 50, 500, or 1,000 μg and the higher value is about 10,000; 20,000; 30,000; or 50,000 μg. Dosage values for a human typically range from about 500 μg to about 50,000 μg per 70 kilogram patient. Boosting dosages of between about 1.0 μg to about 50,000 μg of polypeptide pursuant to a boosting regimen over weeks to months can be administered depending upon the patient's response and condition as determined by measuring the antibody levels or specific activity of CTL and HTL obtained from the patient's blood.

[0223] A human unit dose form of the protein or nucleic acid composition is typically included in a pharmaceutical composition that comprises a human unit dose of an acceptable carrier, typically an aqueous carrier, and is administered in a volume of fluid that is known by those of skill in the art to be used for administration of such compositions to humans (*see*, *e.g.*, Remington "*Pharmaceutical Sciences*", 17 Ed., Gennaro (ed.), Mack Publishing Co., Easton, Pennsylvania (1985)).

[0224] The ancestor or COT proteins and nucleic acids can also be administered via liposomes, which serve to target the peptides to a particular tissue, such as lymphoid tissue, or to target selectively to infected cells, as well as to increase the half-life of the composition. Liposomes include emulsions, foams, micelles, insoluble monolayers, liquid crystals, phospholipid dispersions, lamellar layers and the like. In these preparations, the protein or nucleic acid to be delivered is incorporated as part of a liposome, alone or in conjunction with a molecule that binds to a receptor prevalent among lymphoid cells, such as monoclonal antibodies that bind to the CD45 antigen, or with other therapeutic or immunogenic compositions. Thus, liposomes either filled or decorated with a desired protein or nucleic acid can be directed to the site of lymphoid cells, where the liposomes then deliver the protein compositions to the cells. Liposomes for use in accordance with the invention are formed from standard vesicle-forming lipids, which generally include neutral and negatively

charged phospholipids and a sterol, such as cholesterol. The selection of lipids is generally guided by consideration of, for example, liposome size, acid lability and stability of the liposomes in the blood stream. A variety of methods are available for preparing liposomes, as described in, for example, Szoka *et al.*, *Ann. Rev. Biophys. Bioeng.* 9:467 (1980), and U.S. Patent Nos. 4,235,871; 4,501,728; 4,837,028; and 5,019,369.

[0225] For targeting cells of the immune system, a ligand to be incorporated into the liposome can include, for example, antibodies or fragments thereof specific for cell surface determinants of the desired immune system cells. A liposome suspension containing a protein or nucleic acid can be administered, for example, intravenously, locally, topically, *etc.*, in a dose which varies according to, *inter alia*, the manner of administration, the protein or nucleic acid being delivered, and the like.

[0226] For solid compositions, conventional nontoxic solid carriers can be used which include, for example, pharmaceutical grades of mannitol, lactose, starch, magnesium stearate, sodium saccharin, talcum, cellulose, glucose, sucrose, magnesium carbonate, and the like. For oral administration, a pharmaceutically acceptable nontoxic composition is formed by incorporating any of the normally employed excipients, such as those carriers previously listed, and generally 10-95% of active ingredient, that is, the ancestor or COT proteins or nucleic acids, and typically at a concentration of 25%-75%.

[0227] For aerosol administration, the immunogenic proteins or nucleic acids are typically in finely divided form along with a surfactant and propellant. Suitable percentages of peptides are about 0.01% to about 20% by weight, typically about 1% to about 10%. The surfactant is, of course, nontoxic, and typically soluble in the propellant. Representative of such agents are the esters or partial esters of fatty acids containing from 6 to 22 carbon atoms, such as caproic, octanoic, lauric, palmitic, stearic, linoleic, linolenic, stearic and oleic acids with an aliphatic polyhydric alcohol or its cyclic anhydride. Mixed esters, such as mixed or natural glycerides can be employed. The surfactant can constitute about 0.1% to about 20% by weight of the composition, typically 0.25-5%. The balance of the composition is ordinarily propellant. A carrier can also be included, as desired, as with, for example, lecithin for intranasal delivery.

[0228] *Immune Responses Elicited By The Ancestral or COT Viral Sequences*

[0229] Ancestor or COT proteins (including fragments, derivative and analogs) can be used as a vaccine, as described *supra*. Such vaccines, referred to as a "digital vaccine", are typically screened for those that elicit neutralizing antibody and/or viral (*e.g.*, HIV) specific CTLs against a larger fraction of circulating strains than a vaccine comprising a protein antigen encoded by any sequences of existing viruses or by consensus sequences. Such a digital vaccine will typically provide protection when challenged by the same subtype of virus (*e.g.*, HIV-1 virus) as the subtype from which the ancestral or COT viral sequence was derived.

[0230] The invention also provides methods to analyze the function of ancestral or COT viral gene sequences. For example, in one embodiment, the gp 160 ancestor or COT viral gene sequence is analyzed by assays for functions, such as, for example, CD4 binding, co-receptor binding, receptor specificity (*e.g.*, binding to the CCR5 receptor), protein structure, and the ability to cause cell fusion. Although the ancestor or COT sequences can result in a viable virus, such a viable virus is not necessary for obtaining a successful vaccine. For example, a gpl60 ancestor or COT not correctly folded can be more immunogenic by exposing epitopes that are normally buried to the immune system. Further, although the ancestor or COT viral sequence can be successfully used as a vaccine, such a sequence need not include alternate open reading frames that encode proteins such a tat or rev, when used as an immunogen (*e.g.*, a vaccine).

[0231] Accordingly, in one aspect, mice are immunized with an ancestor or COT protein and tested for humoral and cellular immune responses. Typically, 5-10 mice are intradermally or intramuscularly injected with a plasmid containing a *gag* and/or *env* gene encoding an ancestral or COT viral sequence in, for example, 50 μl volume. Two control groups are typically used to interpret the results. One control group is injected with the same vector containing the *gag* or *env* gene from a standard laboratory strain (*e.g.*, HIV-1-IIIB). A second control group is injected with same vector without any insert. Antibody titration against gag or env protein is performed using standard immunoassays (*e.g.*, ELISA), as described *infra*. The neutralizing antibody is analyzed by subtype-specific laboratory HIV-1 strains, such as for example pNL4-3 (HIV-1-IIIB), as well as primary isolates from HIV-1 infected individuals. The ability of an ancestor or COT viral sequence protein-elicited neutralizing antibody to neutralize a broad primary isolates is one factor indicative of an

immunogenic or vaccine composition. Similar studies can be performed in large animals, such as non-human animals (*e.g.*, macques) or in humans.

[0232] *Immunoassays for titrating the ancestor or COT protein-elicited antibodies*

[0233] There are a variety of assays known to those of ordinary skill in the art for detecting antibodies in a sample (*see*, *e.g.*, Harlow and Lane, *supra*). In general, the presence or absence of antibodies in a subject immunized with an ancestor or COT protein vaccine can be determined by (a) contacting a biological sample obtained from the immunized subject with one or more ancestor or COT proteins (including fragments, derivatives or analogs thereof); (b) detecting in the sample a level of antibody that binds to the ancestor or COT protein(s); and (c) comparing the level of antibody with a predetermined cut-off value.

[0234] In a typical embodiment, the assay involves the use of an ancestor or COT protein (including fragment, derivative or analog) immobilized on a solid support to bind to and remove the antibody from the sample. The bound antibody can then be detected using a detection reagent that contains a reporter group. Suitable detection reagents include antibodies that bind to the antibody/ancestor or COT protein complex and free protein labeled with a reporter group (*e.g.*, in a semi-competitive assay). Alternatively, a competitive assay can be utilized, in which an antibody that binds to the ancestor or COT protein of interest is labeled with a reporter group and allowed to bind to the immobilized antigen after incubation of the antigen with the sample. The extent to which components of the sample inhibit the binding of the labeled antibody to the ancestor or COT protein of interest is indicative of the reactivity of the sample with the immobilized ancestor or COT protein.

[0235] The solid support can be any solid material known to those of ordinary skill in the art to which the antigen may be attached. For example, the solid support can be a test well in a microtiter plate or a nitrocellulose or other suitable membrane. Alternatively, the support can be a bead or disc, such as glass, fiberglass, latex or a plastic material such as polystyrene or polyvinylchloride. The support may also be a magnetic particle or a fiber optic sensor, such as those disclosed, for example, in U.S. Patent No. 5,359,681, the disclosure of which is incorporated by reference herein.

[0236] The ancestor or COT proteins can be bound to the solid support using a variety of techniques known to those of ordinary skill in the art, which are amply described in the patent

and scientific literature. In the context of the present invention, the term "bound" refers to both non-covalent association, such as adsorption, and covalent attachment (*see*, *e.g.*, Pierce *Immunotechnology Catalog and Handbook*, at A12-A13 (1991)).

[0237] In certain embodiments, the assay is an enzyme-linked immunosorbent assay (ELISA). This assay can be performed by first contacting an ancestor or COT protein that has been immobilized on a solid support, commonly the well of a microtiter plate, with the sample, such that antibodies present within the sample that recognize the ancestor or COT protein of interest are allowed to bind to the immobilized protein. Unbound sample is then removed from the immobilized ancestor or COT protein and a detection reagent capable of binding to the immobilized antibody-protein complex is added. The amount of detection reagent that remains bound to the solid support is then determined using a method appropriate for the specific detection reagent.

[0238] More specifically, once the ancestor or COT protein is immobilized on the support as described above, the remaining protein binding sites on the support are typically blocked. Any suitable blocking agent known to those of ordinary skill in the art, such as bovine serum albumin or Tween 20 (Sigma Chemical Co., St. Louis, MO), can be employed. The immobilized ancestor or COT protein is then incubated with the sample, and the antibody is allowed to bind to the protein. The sample can be diluted with a suitable diluent, such as phosphate-buffered saline (PBS) prior to incubation. In general, an appropriate contact time (*i.e.*, incubation time) is a period of time that is sufficient to detect the presence of antibody within a biological sample of an immunized subject. Those of ordinary skill in the art will recognize that the time necessary to achieve equilibrium can be readily determined by assaying the level of binding that occurs over a period of time. At room temperature, an incubation time of about 30 minutes is generally sufficient.

[0239] Unbound sample can then be removed by washing the solid support with an appropriate buffer, such as PBS containing 0.1% Tween 20. Detection reagent can then be added to the solid support. An appropriate detection reagent is any compound that binds to the immobilized antibody-protein complex and that can be detected by any of a variety of means known to those in the art. Typically, the detection reagent contains a binding agent (such as, for example, Protein A, Protein G, immunoglobulin, lectin or free antigen) conjugated to a reporter group. Suitable reporter groups include enzymes (such as

horseradish peroxidase or alkaline phosphatase), substrates, cofactors, inhibitors, dyes, radionuclides, luminescent groups, fluorescent groups, and biotin. The conjugation of a binding agent to the reporter group can be achieved using standard methods known to those of ordinary skill in the art. Common binding agents, pre-conjugated to a variety of reporter groups, can be purchased from many commercial sources (*e.g.*, Zymed Laboratories, San Francisco, CA, and Pierce, Rockford, IL).

[0240] The detection reagent is then incubated with the immobilized antibody- protein complex for an amount of time sufficient to detect the bound antibody. An appropriate amount of time can generally be determined from the manufacturer's instructions or by assaying the level of binding that occurs over a period of time. Unbound detection reagent is then removed and bound detection reagent is detected using the reporter group. The method employed for detecting the reporter group depends upon the nature of the reporter group. For radioactive groups, scintillation counting or autoradiographic methods are generally appropriate. Spectroscopic methods can be used to detect dyes, luminescent groups and fluorescent groups. Biotin can be detected using avidin, coupled to a different reporter group (commonly a radioactive or fluorescent group or an enzyme). Enzyme reporter groups can generally be detected by the addition of substrate (generally for a specific period of time), followed by spectroscopic or other analysis of the reaction products.

[0241] To determine the presence or absence of anti-ancestor or COT protein antibodies in the sample, the signal detected from the reporter group that remains bound to the solid support is generally compared to a signal that corresponds to a predetermined cut-off value. In one embodiment, the cut-off value is the average mean signal obtained when the immobilized ancestor or COT protein is incubated with samples from non-immunized subject.

[0242] In a related embodiment, the assay is performed in a rapid flow-through or strip test format, wherein the ancestor or COT protein is immobilized on a membrane, such as, for example, nitrocellulose, nylon, PVDF, and the like. In the flow-through test, antibodies within the sample bind to the immobilized polypeptide as the sample passes through the membrane. A detection reagent (*e.g.*, protein A-colloidal gold) then binds to the antibody-protein complex as the solution containing the detection reagent flows through the membrane. The detection of bound detection reagent can then be performed as described

above. In the strip test format, one end of the membrane to which the ancestor or COT protein is bound is immersed in a solution containing the sample. The sample migrates along the membrane through a region containing the detection reagent and to the area of immobilized ancestor or COT protein. The concentration of the detection reagent at the protein indicates the presence of anti-ancestor or COT protein antibodies in the sample. Typically, the concentration of detection reagent at that site generates a pattern, such as a line, that can be read visually. The absence of such a pattern indicates a negative result. In general, the amount of protein immobilized on the membrane is selected to generate a visually discernible pattern when the biological sample contains a level of antibodies that would be sufficient to generate a positive signal (*e.g.*, in an ELISA) as discussed *supra*. Typically, the amount of protein immobilized on the membrane ranges from about 25 ng to about 1 μg, and more typically from about 50 ng to about 500 ng. Such tests can typically be performed with a very small amount (*e.g.*, one drop) of subject serum or blood.

[0243] *Cytotoxic T-lymphocyte assay*

[0244] Another factor in treating HIV-1 infection is the cellular immune response, in particular the cellular immune response involving the $CD8^+$ cytotoxic T lymphocytes (CTL's). A cytotoxic T lymphocyte assay can be used to monitor the cellular immune response following sub-genomic immunization with an ancestral or COT viral sequence against homologous and heterologous HIV strains, as above using standard methods (*see*, *e.g.*, Burke *et al.*, *supra*; Tigges *et al.*, *supra*).

[0245] Conventional assays utilized to detect T cell responses include, for example, proliferation assays, lymphokine secretion assays, direct cytotoxicity assays, limiting dilution assays, and the like. For example, antigen-presenting cells that have been incubated with an ancestor or COT protein can be assayed for the ability to induce CTL responses in responder cell populations. Antigen-presenting cells can be cells such as peripheral blood mononuclear cells or dendritic cells. Alternatively, mutant non-human mammalian cell lines that are deficient in their ability to load class I molecules with internally processed peptides and that have been transfected with the appropriate human class I gene, can be used to test the capacity of an ancestor or COT peptide of interest to induce *in vitro* primary CTL responses.

[0246] Peripheral blood mononuclear cells (PBMCs) can be used as the responder cell

source of CTL precursors. The appropriate antigen-presenting cells are incubated with the ancestor or COT protein, after which the protein-loaded antigen-presenting cells are incubated with the responder cell population under optimized culture conditions. Positive CTL activation can be determined by assaying the culture for the presence of CTLs that kill radio-labeled target cells, both specific peptide-pulsed targets as well as target cells expressing endogenously processed forms of the antigen from which the peptide sequence was derived.

[0247] Another suitable method allows direct quantification of antigen-specific T cells by staining with Fluorescein-labeled HLA tetrameric complexes (Altman *et al.*, *Proc. Natl. Acad. Sci. USA* 90:10330 (1993); Altman *et al.*, *Science* 274:94 (1996)). Other relatively recent technical developments include staining for intracellular lymphokines, and interferon release assays or ELISPOT assays. Tetramer staining, intracellular lymphokine staining and ELISPOT assays are typically at least 10-fold more sensitive than more conventional assays (Lalvani *et al.*, *J. Exp. Med.* 186:859 (1997); Dunbar *et al.*, *Curr. Biol.* 8:413 (1998); Murali-Krishna *et al.*, *Immunity* 8:177 (1998)).

[0248] *Diagnosis*

[0249] The present invention also provides methods for diagnosing viral (*e.g.*, HIV) infection and/or AIDS, using the ancestor or COT viral sequences described herein. Diagnosing viral (*e.g.*, HIV) infection and/or AIDS can be carried out using a variety of standard methods well known to those of skill in the art. Such methods include, but are not limited to, immunoassays, as described *supra*, and recombinant DNA methods to detect the presence of nucleic acid sequences. The presence of a viral gene sequence can be detected, for example, by Polymerase Chain Reaction (PCR) using specific primers designed using the sequence, or a portion thereof, set forth in Tables 1 or 3, using standard techniques (*see*, *e.g.*, Innis *et al.*, *PCR Protocols A Guide to Methods and Application* (1990); U.S. Patent Nos. 4,683,202; 4,683,195; and 4,889,818; Gyllensten *et al.*, *Proc. Natl. Acad. Sci. USA* 85:7652-56 (1988); Ochman *et al.*, *Genetics* 120:621-23 (1988); Loh *et al.*, *Science* 243:217-20 (1989)). Alternatively, a viral gene sequence can be detected in a biological sample using hybridization methods with a nucleic acid probe having at least 70% identity to the sequence set forth in Tables 1 or 3, according to methods well known to those of skill in the art (*see*,

e.g., Sambrook *et al.*, *supra*).

[0250] *EXAMPLES*

[0251] *Example 1: Determination of Ancestral Viral Sequences*

[0252] Sequences representing genes of a HIV-1 subtype C were selected from the GenBank and Los Alamos sequence databases. 39 subtype C sequences were used. 18 outgroup sequences (two from each of the other group M subtypes (Figure 8) were used as an outgroup to root the subtype C sequences. The sequences were aligned using CLUSTALW (Thompson *et al.*, *Nucleic Acids Res.* 22:4673-80 (1994)), the alignments were refined using GDE (Smith *et al.*, *CABIOS* 10:671-5 (1994)), and amino acid sequences translated from them. Gaps were manipulated so that they were inserted between codons. This alignment (alignment I) was modified for phylogenetic analysis so that regions that could not be unambiguously aligned were removed (Learn *et al.*, *J. Virol.* 70:5720-30 (1996)) resulting in alignment II.

[0253] An appropriate evolutionary model for phylogeny and ancestral state reconstructions for these sequences (alignment II) was selected using the Akaike Information Criterion (AIC) (Akaike, *IEEE Trans. Autom. Contr.* 19:716-23 (1974)) as implemented in Modeltest 3.0 (Posada and Crandall, Bioinformatics 14: 817-8 (1998)). For the analysis for the subtype C ancestral sequence the optimal model is equal rates for both classes of transitions and different rates for all four classes of transversions, with invariable sites and a X distribution of site-to-site rate variability of variable sites (referred to as a TVM+I+G model). The parameters of the model in this case were: equilibrium nucleotide frequencies: $f_A = 0.3576, f_C = 0.1829, f_G = 0.2314, f_T = 0.2290$; proportion of invariable sites = 0.2447; shape parameter (α) of the X distribution = 0.7623; rate matrix (R) matrix values: $R_{A\rightarrow C} = 1.7502$, $R_{A\rightarrow G} = R_{C\rightarrow T} = 4.1332$, $R_{A\rightarrow T} = 0.6825$, $R_{C\rightarrow G} = 0.6549$, $R_{G\rightarrow T} = 1$.

[0254] Evolutionary trees for the sequences (alignment II) were inferred using maximum likelihood estimation (MLE) methods as implemented in PAUP* version 4.0b (Swofford, PAUP 4.0: Phylogenetic Analysis Using Parsimony (And Other Methods). Sinauer Associates, Inc. (2000)). Specifically for the subtype C sequences, ten different subtree-pruning-regrafting (SPR) heuristic searches were performed each using a different random

addition order. All ten searches found the same MLE phylogeny (LnL = -33585.74). The ancestral nucleotide sequence for subtype C was inferred to be the sequence at the basal node of this subtype using this phylogeny, the sequences from the databases (alignment II), and the TVM+I+G model above using marginal likelihood estimation (see below).

[0255] This inferred sequence does not include predicted ancestral sequence for portions of several variable regions (V1, V2, V4 and V5) and four additional short regions that could not be unambiguously aligned (these eight regions were removed from alignment I to produce alignment II). The following procedure was used to predict amino acid sequences for the complete gp160 including the highly variable regions. The inferred ancestral sequence was visually aligned to alignment I and translated using GDE (Smith *et al.*, *supra*). Since the highly variable regions were deleted as complete codons, the translation was in the correct reading frame and codons were properly maintained. The ancestral amino acid sequence for the regions deleted from alignment II were predicted visually and refined using a parsimony-based sequence reconstruction for these sites using the computer program MacClade, version 3.08a (Maddison and Maddison. MacClade — Analysis of Phylogeny and Character Evolution — Version 3. Sinauer Associates, Inc. (1992)). This amino acid sequences was converted to DNA sequence optimized for expression in human cells using the BACKTRANSLATE program of the Wisconsin Sequence Analysis Package (GCG), version 10 and a human gene codon table from the Codon Usage Database (www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=Homo+sapiens+[gbpri]).

[0256] *Example 2*:

[0257] Different methods are available to determine the maximum likelihood phylogeny for a given subtype. One such method is based on the coalescent theory, which is a mathematical description of the genealogy of a sample of gene sequences drawn from a large evolving population. Coalescence analysis takes into account the HIV population *in vivo* and in the larger epidemic and offers a way of understanding how sampled genealogies behave when different processes operate on the HIV population. This theory can be used to determine the sequence of the ancestral viral sequence, such as a founder, or MRCA. Exponentially growing populations have decreasing coalescent intervals going back in time, while the converse is true for a declining population.

[0258] Epidemics in the USA and Thailand are growing exponentially. The coalescent dates for subtype B epidemics in the USA and Thailand are in accordance with the epidemiologic data. The coalescent date for subtype E epidemic in Thailand is earlier than predicted from the epidemiologic data. Potential reasons that can account for this discrepancy include, for example, the existence of multiple introductions of HIV-1 (there is no evidence from phylogenetics on this point), the absence of HIV-1 detection in Thailand for about 7 years, and the difference in the mutation rates for env gene in the HIV-1 subtypes E and B.

[0259] *The unit of reconstruction*

[0260] This unit of reconstruction relates to the ancestral viral sequence (*i.e.*, state) state that is reconstructed. There are three possible units of reconstruction: nucleotides, amino acids or codons. In one embodiment, the states of the individual nucleotides are reconstructed and the amino acid sequences are then determined on the basis of this reconstruction. In another embodiment, the amino acid ancestral states are directly reconstructed. In a typical embodiment, the codons are reconstructed using a likelihood-based procedure that uses a codon model of evolution. A codon model of evolution takes into account the frequencies of the codons and implicitly the probability of substituting one nucleotide for another - in other words, it incorporates both nucleotide and amino acid substitutions in a single model. Computer programs capable of doing this are available or can readily be developed, as will be appreciated by the skilled artisan.

[0261] *Use of marginal or joint likelihoods for estimating the ancestral states*

[0262] The ancestral state can be estimated using either a marginal or a joint likelihood. The marginal and joint likelihoods differ on the basis of how ancestral states at other nodes in the phylogenetic tree estimated. For any particular tree, the probability that the ancestral state of a given site on a sequence alignment at the root is, for example, an A can be determined in different ways.

[0263] The likelihood that the nucleotide is an adenine (A) can be determined regardless of whether higher nodes (i.e., those nodes closer to the ancestral viral sequence, founder or MRCA) have an adenine, cytosine (C), guanine(G), or thymine (T). This is the marginal likelihood of the ancestral state being A.

[0264] Alternatively, the likelihood that the nucleotide is an A can be determined depending on whether the nodes above are A, C, G, or T. This estimation is the joint likelihood of A with all the other ancestral reconstructions for that site.

[0265] The joint likelihood is a preferred method when all the ancestral states along the entire tree need to be determined. To establish the most likely states at one given node, the marginal likelihood is preferably used. In case of uncertainty at a particular site, a likelihood estimate of the ancestral state allows testing whether one state is statistically better than another. If two possible ancestral states do not have statistically different likelihoods, or if one ends up with multiple states over a number of sites building all possible sequences is not desirable. The likelihoods of all combinations can however be computed and ranked, and only those above a certain critical value are used. For example, when two sites on a sequence, each with different likelihoods for A, C, G, T, are considered:

L(A) L(C) L(G) L(T)* * L represents the -lnL (the negative log-likelihood); therefore, the smaller the more likely.

Site 1 3 2 1.5 1

Site 2 10 7 5 1

there are 16 possible sequence configurations, each with its own log-likelihood, that is simply the sum of the log-likelihoods for each base, which are:

AA 13	CA12	GA 11.5	TA 11
AC 10	CC 9	GC 8.5	TC 8
AG 8	CG 7	GG 6.5	TG 6
AT 4	CT 3	GT 2.5	TT 2

[0266] In order of likelihood the ranking is:

TT, GT, CT, AT, TG, GG, CG, AG, TC, GC, CC, AC, TA, GA, CA, AA

[0267] The first four sequences have T at the second site. This results from the likelihood at that site being spread over a large range, resulting into a very low probability of having any nucleotide other than T at this site. At Site 1, however, any nucleotide tends to give quite similar likelihoods. This kind of ranking is one way of whittling down the number of

possible sequences to look at if variation is to be taken into account.

[0268] The above variation in reconstructed ancestral states deals with variation that comes about because of the stochastic nature of the evolutionary process, and because of the probabilistic models of that process that are typically used. Another source of variation results from the sampling of sequences. One way of testing how sampling affects ancestral state reconstruction is to perform jackknife re-sampling on an existing data set. This involves deleting randomly without replacement of some portion (*e.g.*, half) of the sequences, and reconstructing the ancestral state. Alternatively, the ancestral state can be estimated for each of a set of bootstrap trees, and the number of times a particular nucleotide was estimated can be reported as the ancestral state for a given site. The bootstrap trees are generated using bootstrapped data, but the ancestral state reconstructions use the bootstrap trees on the original data.

[0269] Different models of evolution can be used to reconstruct the ancestral states for the root node. Examples of models are known and can be chosen on a multitude of levels. For example, a model of evolution can be chosen by some heuristic means or by picking one that gives the highest likelihood for the ancestral sequence (obtained by summing the likelihoods over all sites). Alternatively the ancestral states are reconstructed at each site over all models of evolution, all of the likelihoods obtained summed, and the ancestral state chosen that has the maximum likelihood.

[0270] *Example 3*:

[0271] The conservation of HIV-1 subtype C CTL amino acid consensus epitopes was analyzed. The total number of epitopes was 395. The table below summarize the results of the similarly of each circulating viral sequence to the C subtype CTL consensus sequence. The determined ancestor viral sequence for the HIV-1 subtype C env protein (SEQ ID NO:4) has the highest score (98.48%). Note that the scores for several strains are below 65%, because truncated sequences were used.

Sequence Name	Total AA number	Percentage CTL to Consensus
cCanc95-mod1	389	98.48%
cBR.92BR025	376	95.19%
cBI.BU910717	363	91.90%

	cIN.21068	368	93.16%
	cIN.301905	370	93.67%
	cMW959.U08453	358	90.63%
	cBW.96BW1210	365	92.41%
5	cBI.BU910316	367	92.91%
	cZAM176.U86778	352	89.11%
	cMW965.U08455	364	92.15%
	cZAM174.16.U86768	351	88.86%
	c84ZR085.U88822	322	81.52%
10	cSN.SE364A	370	93.67%
	cMW960.U08454	365	92.41%
	cBI.BU910812	368	93.16%
	cET.ETH2220	358	90.63%
	cBI.BU910518	361	91.39%
15	cIN.94IN11246	361	91.39%
	cBW.96BW15B03	359	90.89%
	cDJ.DJ259A	355	89.87%
	cBI.BU910213	365	92.41%
	cBW.96BW01B03	362	91.65%
20	cIND760.L07655	255	64.56%
	cIN.301904	372	94.18%
	cSO.SM145A	354	89.62%
	cCHN19.AF268277	356	90.13%
	cIND747.L07653	255	64.56%
25	cBW.96BW0402	364	92.15%
	cBI.BU910611	367	92.91%
	cBI.BU910423	359	90.89%
	cBW.96BW17B05	355	89.87%
	cBW.96BW0502	367	92.91%
30	cUG.UG268A2	372	94.18%
	cZAM18.L22954	365	92.41%
	cIN.301999	368	93.16%

c91BR15.U39238	371	93.92%
cDJ.DJ373A	361	91.39%
cBI.BU910112	369	93.42%
c93IN101.AB023804	365	92.41%
cBW.96BW16B01	361	91.39%
cBW.96BW11B01	361	91.39%
cINdiananc66	363	91.90%

[0272] *Example 4*:

[0273] Ancestor sequence reconstruction was performed on simian immunodeficiency viruses grown in macaques. Macaques were infected and challenged with a relatively homogeneous SIV inoculum. Viral sequences were obtained up to three years following infection and were used to deduce an MRCA using maximum likelihood phylogeny analysis. The resulting sequence was compared to the consensus sequence of the inoculum. The MRCA sequence was found to be 97.4% identical to the virus inoculum. This figure improved to 98.2% when convergence at 5 glycosylation sites was removed - this convergence was due to readaptation of the virus from tissue culture to growth in the animal (Edmonson *et al.*, *J. Virol.* 72:405-14 (1998)). The MRCA sequence and the consensus sequence were found to differ at 1.5% at the nucleotide level. Figure 3 illustrates the determination of simian immunodeficiency virus MRCA phylogeny.

[0274] *Example 5*:

[0275] An experiment to test the biological activity of the HIV-1 subtype B ancestral viral env gene sequence was performed. A nucleic acid sequence encoding the HIV-1 subtype B ancestral viral env gene sequence was assembled from long (160-200 base) oligonucleotides; the assembled gene was designated ANC1. The biological activity of ANC1 HIV-1-B Env was evaluated in co-receptor binding and syncytium formation assays. The plasmid pANC1, harboring the determined and chemically synthesized HIV-1 subtype B Ancestor gp160 Env sequence, or a positive control plasmid containing the HIV-1 subtype B 89.6 gp160 Env, was transfected into COS7 cells. These cells are capable of taking up and expressing foreign DNA at high efficiencies and thus are routinely used to produce viral proteins for

presentation to other cells. The transfected COS7 cells were then mixed with GHOST cells expressing either one of the two major HIV-1 co-receptor proteins, CCR5 or CXCR4. CCR5 is the predominant receptor used by HIV early in infection. CXCR4 is used later in infection, and use of the latter receptor is temporally associated with the development of disease. The COS7-GHOST-co-receptor+ cells were then monitored for giant cell formation by light microscopy and for expression of viral Env protein by HIV-Env-specific antibody staining and fluorescence detection.

[0276] Cells expressing the ANC1 Env were shown to be expressed by virtue of binding to HIV-specific antibody and fluorescent detection, and to cause the formation of giant multinucleated cells in the presence of the CCR5 co-receptor, but not the CXCR4 co-receptor. The positive control 89.6 Env uses both CCR5 and CXCR4 and formed syncytia with cells expressing either co-receptor. Thus, the ANC1 Env protein was shown to be biologically active by co-receptor binding and syncytium formation.

[0277] *Example 6*:

[0278] Maximum likelihood phylogeny reconstruction differs from traditional consensus sequence determinations because a consensus sequence represents a sequence of the most common nucleotide or amino acid residue at each site in the sequence. Thus, a consensus sequence is subject to biased sampling. In particular, the determination of a consensus sequence can be biased if many samples have the same sequence. In addition, the consensus sequence is a real viral sequence.

[0279] In contrast, maximum likelihood phylogeny analysis is less likely to be affected by biased sample because it does not determine the sequence of a most recent common ancestor based solely on the frequencies of the each nucleotide at each position. The determined ancestral viral sequence is an estimate of a real virus, the virus that is the common ancestor of the sampled circulating viruses.

[0280] In the simplest of methods for determining an ancestral sequence, for a single site on a sequence alignment nucleotides are assigned to ancestral nodes such that the total number of changes between nodes is minimized; this approach is called a "most parsimonious reconstruction." An alternative methodology, based on the principle of maximum likelihood, assigns nucleotides at the nodes such that the probability of obtaining

the observed sequences, given a phylogeny, is maximized. The phylogeny is constructed by using a model of evolution that specifies the probabilities of nucleotide substitutions. The maximum likelihood phylogeny is the one that has the highest probability of giving the observed data.

[0281] Referring to Figure 5, a comparison is presented of parsimony methodology and maximum likelihood methodology of determining an ancestral viral sequence (*e.g.*, a founder sequence or a most recent common ancestor sequence (MRCA)). The most parsimonious reconstruction ("MP") can have the undesirable problem of creating an ambiguous state at the ancestral branch point (*i.e.*, node). In this example, the two descendant sequences from this node have an adenine (A) or guanine (G) at a particular position in the sequence. The most parsimonious reconstruction ("MP Reconstruction") for the ancestral sequence at this site is ambiguous, because there can be either an A or G (symbolized by "R") at this position. In contrast, a maximum likelihood phylogeny analysis applies knowledge about sequence evolution. For example, likelihood analysis relies, in part, on the identity of nucleotides at the same position in other variants. Thus, in this example, a G to A mutation is more likely than an A to G change because variant at the adjacent node also has a G at the same position.

[0282] Referring to Figure 6, another example illustrates the differences in these methodologies to determine a most recent common ancestor. In this example, twelve sequences of seven nucleotides are presented. These sequences share the illustrated evolutionary history. A consensus sequence calculated from these sequences is CATACTG. In panel A, the maximum likelihood reconstruction of the determined ancestral node is shown as GATCCTG. Other determined sequences are presented adjacent the other internal nodes. In panel B, the most parsimonious reconstruction at the same nodes is presented. As shown, the most parsimonious reconstruction predicts the consensus sequence GAWCCTG, where "W" symbolizes that either an A or T is equally possible to be at the third position. Similarly other most parsimonious reconstructions are shown at the various internal nodes. At the seventh internal node, the last nucleotide is indicated with the symbol "V" representing that an A, C or G might be present. Also note in this example, the consensus sequence differs in at least two sites (the 1st and 4th positions) from either the maximum likelihood- or parsimony-determined sequence for the MRCA.

[0283] *Example 7*:

[0284] MRCA and COT state reconstructions were performed. Complete HIV-1 genome sequences were obtained from the Los Alamos National Laboratory HIV Sequence database or from research done in support of the HIV Vaccine Trials Network (HVTN) in the Mullins laboratory (HVTN1925c1.US98, HVTN3605c9.US98, HVTN8229c6.US98, and HVTN941c16.US98). Each sequence set, subtype B (Table 7) and subtype C (Table 9), was aligned using CLUSTALW (Thompson, 1994 #6159). Resulting alignments were examined and adjusted using amino acid alignments as guides with MacClade version 4 (Maddison, 2001 #13641). Alignment gaps were inserted between adjacent codons. Regions that could not be unambiguously aligned were omitted from the following phylogenetic analyses but included in the ancestral state reconstructions (see below). The complete genome alignments were partitioned into 9 gene coding regions (*gag*, *env* (encoding gp160), *nef*, *pol*, *rev*, *tat*, *vif*, *vpr*, and *vpu*); *tat* and *rev* exon sequences were spliced together after the intervening intron sequence was removed to produce a single continuous protein coding region.

[0285] Parameter values of optimal evolutionary models for each of these nucleotide alignments, except for subtype C env and gag, were estimated using MODELTEST (Posada, 1998 #10104) in conjunction with PAUP* (Swofford, 1999 #9501) based on the Akaike Information Criterion (Akaike, 1974 #10404). For subtype C env and gag, PAUP* was used to estimate HKY parameter values (Hasegawa, 1985 #9867) based on a neighbor-joining (Saitou, 1987 #5332) phylogenetic tree. Evolutionary model parameter values are presented for the subtype B genes (Table 8) and the subtype C genes (Table 10). These values were used for maximum-likelihood phylogenetic analyses using PAUP* (Swofford, 1999 #9501). For all genes except subtype C *env* and *gag*, 10 random-addition subtree-pruning-regrafting iterations were done. For subtype C *env* and *gag*, a single iteration based on a neighbor-joining starting trees were used.

[0286] After phylogenetic trees were obtained, the regions of ambiguous alignment that were removed (see above) were included in their original positions in the alignments. For the MRCA estimation, the sequence states at the ancestral node (the point at which the subtype D sequences attach to the portion of the tree that solely included subtype B sequences) were

derived using marginal maximum-likelihood estimation with PAUP* {Swofford, 1999 #9501}. The MRCA protein sequence is the translation of the MRCA nucleotide sequence. For the COT state reconstructions, the outgroup sequences (subtype D sequences in Table 7; subtype other than C in Table 9) were pruned from the phylogenetic tree for each gene using PAUP* (Swofford, 1999 #9501). The remaining gene trees were analyzed using perl computer programs to determine the point at the center of the phylogenetic tree via the Minimum of Means, and Least-Squares Method. Evolutionary states at this point for each method for each gene tree were then estimated using marginal state estimation using PAUP* (Swofford, 1999 #9501).

Table 7: Sequence Names, GenBank Accession Numbers and Country of Isolation for the sequences used in the complete genome MRCA and COT estimation of subtype B. §: subtype D sequences included as an outgroup to root the clade B phylogeny; *: unpublished complete genome sequences from HVTN project.

Sequence	Accession number	Country	Subtype§
1WK.KR97	AF224507	Korea	B
3202A21.NL86	U34604	Netherlands	B
89SP061.ES89	AJ006287	Spain	B
AD8.US86	AF004394	U.S.A.	B
ARCH054.AR98	AY037268	Argentina	B
ARMA173.AR99	AY037274	Argentina	B
ARMS008.AR00	AY037269	Argentina	B
BK132.TH90	AY173951	Thailand	B
BOL122.BO99	AY037270	Bolivia	B
BZ167.BR89	AY173956	Brazil	B
CAM1.GB83	D10112	United Kingdom	B
D31.DE86	U43096	Germany	B
DH123.US91	AF069140	U.S.A.	B
HAN.DE86	U43141	Germany	B
HVTN1925c1.US98	*	U.S.A.	B
HVTN3605c9.US98	*	U.S.A.	B
HVTN8229c6.US98	*	U.S.A.	B
HVTN941c16.US98	*	U.S.A.	B
HXB2.FR83	K03455	France	B
JRCSF.US86	M38429	U.S.A.	B
MBC200.AU86	AF042100	Australia	B
MBC925.AU87	AF042101	Australia	B
MBCC18B.AU93	AF042102	Australia	B
MBCC98.AU96	AF042104	Australia	B
MBCD36.AU96	AF042105	Australia	B
MN.US84	M17449	U.S.A.	B
NY5.US84	M38431	U.S.A.	B
OYI.GA88	M26727	Gabon	B

P896.US89	U39362	U.S.A.	B
RF.US83	M17451	U.S.A.	B
RL42.CN	U71182	China	B
SF2.US83	K02007	U.S.A.	B
TWCYS.TW94	AF086817	Taiwan	B
WCIPR9018.US90	U69591	U.S.A.	B
WEAU160.US90	U21135	U.S.A.	B
WR27.US88	U26546	U.S.A.	B
YU2.US86	M93258	U.S.A.	B
D.ELI.CD83	K03454	D. R. Congo	D
D.UG1141.UG94	U88824	Uganda	D
D.ZR085.CD84	U88822	D. R. Congo	D

Table 8: Evolutionary model parameters for phylogenetic analysis and ancestral and COT state reconstruction for subtype B gene sequences.

gene	$R_{A<->C}$	$R_{A<->G}$	$R_{A<->T}$	$R_{C<->G}$	$R_{C<->T}$	$R_{G<->T}$	Shape parameter	Assumed proportion of invariable sites	freq(A)	freq(C)	freq(G)	freq(T)
gag	1.787	4.4321	0.8198	0.9897	5.2536	1	0.5253	0.3028	0.3832	0.1751	0.2268	0.2149
gp160	2.469	5.6546	1.1659	0.9752	5.6546	1	0.5833	0.2469	0.3575	0.1858	0.2236	0.2331
nef	1.2074	3.3345	1.193	0.5574	3.3345	1	0.5329	none	0.346	0.2094	0.2613	0.1833
pol	2.6829	11.9199	1.1681	1.342	11.9199	1	0.6476	0.369	0.3937	0.1717	0.2138	0.2208
rev	1	2.1615	0.3786	0.3786	2.1615	1	0.4075	none	0.3224	0.2751	0.2411	0.1614
tat	1.9279	3.0288	0.3625	1.2287	3.0288	1	0.5314	0.1979	0.3543	0.2489	0.2191	0.1777
vif	1.7739	3.7387	0.3906	0.5473	3.7387	1	0.5248	0.2459	0.3796	0.1809	0.2316	0.2079
vpr	3.657	7.8789	1.0026	1.3814	15.7753	1	0.6295	0.3193	0.3439	0.1894	0.2386	0.2281
vpu	2.3344	3.1513	1.4396	2.2202	8.3028	1	0.5218	none	0.368	0.1192	0.26	0.2528

Table 10: Evolutionary model parameters for phylogenetic analysis and ancestral and COT state reconstruction for subtype C gene sequences.

gene	$R_{A<->C}$	$R_{A<->G}$	$R_{A<->T}$	$R_{C<->G}$	$R_{C<->T}$	$R_{G<->T}$	Shape parameter	Assumed proportion of invariable sites	freq(A)	freq(C)	freq(G)	freq(T)	Ti/Tv
gag	-	-	-	-	-	-	0.3348	none	0.36904	0.18619	0.24234	0.20244	3.09542
gp160	-	-	-	-	-	-	0.405447	none	0.34797	0.17109	0.23644	0.24449	2.3953
nef	2.0998	4.0826	1.4418	1.3501	5.7703	1	0.6877	0.1885	0.3426	0.2063	0.2579	0.1932	-
pol	1.9799	8.7785	1.078	0.7924	11.9393	1	0.7828	0.3661	0.3961	0.1705	0.2231	0.2103	-
rev	2.1477	5.7666	0.7273	0.3826	5.7666	1	0.71	0.1773	0.3116	0.2212	0.2668	0.2004	-
tat	4.1494	7.2595	1.082	1.0846	9.5448	1	0.6771	0.3018	0.3463	0.2367	0.2339	0.1831	-
vif	1.6981	4.1865	0.8836	0.7624	5.8585	1	0.525	0.2597	0.3842	0.1794	0.2363	0.2001	-
vpr	2.571	6.3751	0.9793	0.8462	9.8567	1	0.829	0.3024	0.3448	0.1814	0.2495	0.2243	-
vpu	1.5454	2.1322	0.5312	0.2609	5.7821	1	0.635	0.1585	0.4191	0.1036	0.2453	0.232	-

Table 9: Sequence Names, GenBank Accession Numbers and Country of Isolation for the sequences used in the complete genome MRCA and COT estimation of subtype C. §: sequences from subtypes other than D included as an outgroup to root the clade C phylogeny.

5

Sequence	Accession number	Country	SUBTYPE§
86ETH2220	U46016	Ethiopia	C
92BR025	U52953	Brazil	C
93IN904	AF067157	India	C
94IN476	AF286223	India	C
95IN21068	AF067155	India	C
96BW0402	AF110962	Botswana	C
96BW0408	AF110964	Botswana	C
96BW0502	AF110967	Botswana	C
96BW06J4	AF290028	Botswana	C
96BW1106	AF110970	Botswana	C
96BW1210	AF110972	Botswana	C
96BW15B03	AF110973	Botswana	C
96BW16B01	AF110976	Botswana	C
96BW17A09	AF110979	Botswana	C
96BW96BW01B03	AF110959	Botswana	C
96BWM032	AF443075	Botswana	C
96ZM651	AF286224	Zambia	C
96ZM751	AF286225	Zambia	C
97TZ04	AF361874	Tanzania	C
97TZ05	AF361875	Tanzania	C
97ZA012	AF286227	South Africa	C
98BR004	AF286228	Brazil	C
98BWMO1410	AF443079	Botswana	C
98BWMO18D5	AF443080	Botswana	C
98BWMO36A5	AF443081	Botswana	C
98BWMO37D5	AF443082	Botswana	C
98IN012	AF286231	India	C
98IS002	AF286233	Israel	C
98TZ013	AF286234	Tanzania	C
98TZ017	AF286235	Tanzania	C
99BW393212	AF443083	Botswana	C
99BW46424	AF443084	Botswana	C
99BW47458	AF443085	Botswana	C
99BW47547	AF443086	Botswana	C
00BW076820	AF443089	Botswana	C
00BW087421	AF443090	Botswana	C
00BW147127	AF443091	Botswana	C
00BW16162	AF443092	Botswana	C
00BW1686.	AF443093	Botswana	C
00BW17593	AF443094	Botswana	C
00BW17835	AF443096	Botswana	C
00BW17956	AF443097	Botswana	C
00BW18113	AF443098	Botswana	C
00BW18802	AF443100	Botswana	C
00BW22767	AF443107	Botswana	C

00BW38713	AF443110	Botswana	C
00BW38769	AF443111	Botswana	C
00BW38868	AF443112	Botswana	C
00BW50311	AF443115	Botswana	C
ZA. .CTSC2	AY043176	South Africa	C
ZA. .DU151	AY043173	South Africa	C
ZA. .DU179	AY043174	South Africa	C
A KE.Q2317	AF004885	Kenya	A
B US.JRFL	U63632	U.S.A.	B
D UG.94UG1	U88824	Uganda	D
F BE.VI850	AF077336	Belgium	F
G SE.SE616	AF061642	Sweden	G
J SE.SE928	AF082394	Sweden	J
H CF.90CF0	AF005496	C.A.R.	H

[0287] Comparisons of the MRCA, Least Squares Method ("LSCOT") and Minimum of Means COT ("MMCOT") reconstructions for the Clade B *gag*, *env* (encoding gp160), *nef*, *pol*, *rev*, *tat*, *vif*, *vpr*, and *vpu* genes are shown in Figures 9 to 17. Comparisons of the MRCA, Least Squares Method ("LSCOT") and Minimum of Means COT ("MMCOT") reconstructions for the Clade B gag, env (gp160), nef, pol, rev, tat, vif, vpr, and vpu proteins are shown in Figures 18 to 26.

[0288] Comparisons of the MRCA, Least Squares Method ("LSCOT") and Minimum of Means COT ("MMCOT") reconstructions for the Clade C *gag*, *env* (encoding gp160), *nef*, *pol*, *rev*, *tat*, *vif*, *vpr*, and *vpu* genes are shown in Figures 27 to 35. Comparisons of the MRCA, Least Squares Method ("LSCOT") and Minimum of Means COT ("MMCOT") reconstructions for the Clade C gag, env (gp160), nef, pol, rev, tat, vif, vpr, and vpu proteins are shown in Figures 36 to 44.

Example 8:

[0289] HIV-1 has high replication and mutation rates that permit rapid generation of viruses that can escape immune recognition. Within an infected host, the HIV-1 population diversifies over time, producing mostly defective viruses but nonetheless persisting and accumulating mutations at a rate of up to 1% per year in its *env* gene. HIV sequences sampled from a population of infected individuals recapitulate a star-like phylogeny, *i.e.*, most of the variants sampled at the same time are positioned roughly equidistant from the center of the tree. Thus, any given variant is approximately twice this distance from any other circulating strain. A primary concern in designing protective AIDS vaccines, then, is

the choice of strains likely to best provide protection against the expanding population of HIV-1 variants.

[0290] Several methods for choosing a vaccine candidate on the basis of genetic or protein sequence data have been put forth recently. First, and the approach followed in current clinical trials, is to choose one or a small number of laboratory-grown or primary viral isolates, typically chosen to approximate a "circulating" strain or to simply match the HIV-1 subtype(s) in the targeted population. An advantage of this approach is that it typically employs viral genes derived from a viable virus and thus produces antigens likely to adopt "native" conformations. However, as a result of HIV-1 mutational radiation, any given "circulating" strain will be genetically, and presumably antigenically, dissimilar to other strains likely to be encountered by the vaccinee, with the degree of dissimilarity proportional to the length of time the virus has been circulating within the population. Thus, vaccines based on specific viral isolates are unlikely to be effective against a broad range of circulating viruses. The results of a first Phase III AIDS vaccine trial suggest that monomeric envelope proteins that are derived from such isolates are insufficient to provide protective immunity, although it remains an open question whether more native presentations of envelope protein might be an effective vaccine component.

[0291] To enhance the breadth of the elicited immune response, a second approach is to include as many diverse HIV-1 isolates as possible in the vaccine recipe, with the intention of inducing multiple responses against different Env proteins. This approach has been explored to a limited degree, without clear success. A third approach is to build a consensus sequence based on either circulating strains or strains in the HIV database. The consensus sequence will be genetically closer to circulating strains than any given natural virus isolate, but its sequence may be biased by sampling and may link polymorphisms in combinations not found in any natural or viable virus, thus potentially resulting in inappropriate structural conformations. Consequently, there is a need for new, effective methods of identifying candidate sequences for vaccine development to treat and/or prevent HIV infection.

[0292] To this end, the use of an HIV population ancestral sequence can be used as a vaccine candidate. Such a vaccine might correspond to an ancestor of all known HIV strains, an HIV sequence subtype, or viruses circulating in a given geographic region or risk group. The ancestral viral sequence is reconstructed from a phylogenetic tree describing the

historical relationships of sequences sampled from the population of interest, and is thus expected to correspond to the most recent common ancestor (MRCA) of the viral strains sampled from the targeted population. It is also likely that such an ancestral sequence will encompass elements conserved within the sampled virus population. To maximize the opportunity for native Envelope protein expression *in vivo*, DNA vaccines expressing full-length Env gp160 or TM-truncated Env gp140 have been used. These vaccines have been shown are effective in both priming and boosting in macaques. Here, a predicted ancestral *env* sequence for subtype B HIV-1 encodes a functional protein that elicits neutralizing antibodies to primary isolates in rabbits. Boosting with gp120 protein did not significantly enhance the neutralization capacity of the sera.

Methods

[0293] *Ancestral state simulations.* To assess the accuracy of the ancestral state reconstruction method, the following *in silico* study was performed. Using the program Seq-Gen (Rambaut and Grassly, *Comput. Appl. Biosci.* 13:235-8 (1997)), the evolution of sequences was simulated, given an ancestral sequence and following the phylogenetic histories depicted in the trees in Figure 45A and the model of evolution shown in Table 11. One hundred replicate data sets were generated for both tree shapes. The PAUP* (Swofford, D. 1999. PAUP*, Sinauer Associates, Inc.) to generate a maximum-likelihood tree and an ancestral sequence for all 200 datasets (100 simulations on both the star- and caterpillar-shaped phylogenies). To be conservative, the simulations were performed on trees with very long external branches (having a genetic distance of 20%), and the caterpillar trees had internal branch lengths of 1% genetic distance.

[0294] *Ancestral state reconstruction of the SIVmacBK28 env sequence*: SIVmac sequences obtained from a series of experimental infections of rhesus macaques (Edmonson *et al.*, *J. Virol.* 72:405-14 (1998)) were used to reconstruct an ancestral sequence to compare to the plasmid-derived SIV clone used to inoculate these monkeys. A region of *env* from position 8,265 to 8,827 of the SIVmacBK28 genome was amplified by nested-PCR from uncultured PBMC DNA from the SIV infected macaques. First round primers were UP-3, positions (8,117 to 8,130), AGACTGCAGATGTGAAGAGGTACAC and PEXTM6 (8,977 to 8,953), GGATCTGGTATGCTCATAGCAA. Second round primers were PEXTM7 (8,265 to 8,286), GATACTGCAGCAACAGCAACAGCTG and UP-5 (8,827 to 8,810),

GCAAAGCTTCTCTGGTTGGCAGTG. Amplified products were then cloned and sequenced. The GenBank accession numbers for these sequences are AY169007–AY169163. Methods for SIVmacBK28 ancestor reconstruction were identical to those outlined below. The model of evolution is given in Table 12 and the reconstructed sequence is provided in Figure 46.

[0295] *Ancestral state reconstruction of an HIV-1-B env sequence*: Thirty-eight sequences representing *env* genes of HIV-1 subtype B were selected from GenBank and the Los Alamos sequence databases, with three additional sequences from clade D used as an outgroup for rooting (Kuiken *et al., HIV Sequence Compendium* 2001, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.) (Table 11 and Figure 45B). Sequences were aligned using CLUSTALW (Thompson *et al.*, *Nucleic Acids Res.* 22:4673-80 (1994) and the alignment was refined using GDE (Smith *et al.*, *Comput Appl. Biosci.* 10:671-5 (1994)). Inferred amino acid sequences were used to guide the introduction of alignment gaps such that they were inserted between codons. These alignments were then modified for phylogenetic analysis such that regions that could not be unambiguously aligned were removed (Learn *et al.*, *J. Virol.* 70:5720-30 (1996)).

[0296] An appropriate evolutionary model for phylogeny and ancestral state reconstructions was selected using the Akaike Information Criterion (AIC) as implemented in Modeltest 3.0 (Posada and Crandall, *Bioinformatics* 14:817-8 (1998)). (*See* Table 12). Evolutionary trees for the sequences (B alignments) were inferred using maximum-likelihood estimation methods as implemented in PAUP* version 4.0b10 (Swofford, D. 1999. PAUP*, Sinauer Associates, Inc.). Ten different subtree-pruning-regrafting (SPR) heuristic searches were performed using a different random-addition order each time and using a neighbor-joining tree of the subtype B sequences as a backbone constraint. The ancestral nucleotide sequence was inferred as the sequence at the basal node of the clade using the inferred phylogenies. The nucleotide at each site of the ancestral sequence had the highest likelihood, integrated over all state assignments at other nodes. This derived sequence is provided in Fig. 46.

[0297] To predict the amino acid sequences for the complete gp160, including the highly variable regions, the inferred ancestral sequences were visually aligned to the alignments prior to gap-stripping and translated using GDE (Smith *et al.*, *Comput. Appl. Biosci.* 10:671-5

(1994)). Since the highly variable regions were deleted as complete codon triplets, the translations into amino acids were in the correct reading frame and codons were properly maintained. The ancestral amino acid sequences for the regions deleted from the gap-striped alignment were predicted visually and refined using parsimony-based sequence reconstruction for these sites using the program MacClade (Maddison and Maddison, MacClade - Analysis of Phylogeny and Character Evolution, 3rd ed., Sinauer Associates, Inc, Sunderland, MA. (1992)). The complete gp160 amino acid sequences were converted to DNA sequences optimized for expression in human cells using the BACKTRANSLATE program of the Wisconsin Sequence Analysis Package (GCG), version 10 and a human gene codon table from a Codon Usage Database (Nakamura *et al.*, *Nucleic Acids Res.* 28:292 (2000)). The gene was then chemically synthesized, sequence-verified, and inserted into a mammalian expression vector to produce the plasmid pVR1012-AN1env (Vical, Inc). To generate a gp140 expression plasmid, pVR1012-AN1env was subjected to site-directed mutagenesis (Quik-change kit, Stratagene) to introduce two stop codons after the codon for amino acid 711 (nucleotides 2134-9)

[0298] *Fusion assay.* GHOST-CCR5 or GHOST-CXCR4 cells were seeded in two-chamber culture slides (Falcon, Franklin Lakes, NJ) at $5x10^4$ cells/well, and transfected the next day with 0.5 μg *env* plasmid using Fugene-6 (Roche, Indianapolis, IN). Transfected cells were allowed to fuse for two days and then were fixed and stained for immunofluorescence using the monoclonal antibody b12 (gift of Dr. Dennis Burton) and 1:2500 FITC-conjugated goat anti-human IgG (Cappel). The presence of large multinucleated cells indicated that cell-cell fusion had occurred.

[0299] *Western blot analysis.* 293T cells were transfected in 6-well dishes with 1 μg env plasmid and Fugene-6, and medium and cell lysates were collected 48 hours later. Twenty one μl of each sample were used for SDS-PAGE, transferred to a PVDF membrane, probed with heat-inactivated pooled human HIV^+ serum (gift of Dr. Lisa M. Frenkel) and horseradish peroxidase-conjugated goat anti-human IgG (Sigma, St Louis, MO), and then developed with the color substrate 3-amino-9-ethylcarbazole (AEC).

[0300] *Recombinant protein production.* Recombinant gp120 from HIV-1 SF162 was provided by Chiron Corp and prepared as described in (Barnett *et al.*, *Vaccine* 15:869-73 (1997)).

[0301] *Rabbit immunizations.* New Zealand White rabbits were vaccinated by Gene Gun (Bio-Rad, Hercules, CA). Eighteen shots of 2 μg DNA each were given for each dose. Animals were vaccinated at weeks 1, 5, 10, 19, 23, and 46. Protein boosts were given at weeks 68 and 78; these consisted of 44 μg recombinant gp120 mixed with an equal volume of Incomplete Freund's adjuvant, injected intramuscularly. Blood was collected the day before and 2 weeks after each immunization. Animals were housed at R+R Rabbitry, Marysville, WA and procedures followed IACUC-approved protocols.

[0302] *Neutralization assays. cMAGI assay.* Assays were performed as described by (Doria-Rose *et al.*, *J. Virol.* 77:11563-77 (2003). Briefly, serial dilutions of sera were incubated with virus for one hour, then added to duplicate wells of cMAGI cells. After two days, cells were fixed and stained for ß-galactosidase expression (indicating infection). The percent neutralization at a given titer is calculated by the equation (Vo-Vn)/Vo x 100, where Vn is the number of infected cells in the virus+antibody wells and Vo is the number of positive cells in virus-alone wells. Titers were normalized to the titer of a standard HIV^+ human serum pool that was included on each assay plate.

[0303] *Luciferase reduction assay (M7-luc assay).* 500 $TCID_{50}$ of virus was mixed and incubated with serial dilutions of serum in triplicate for 1 hr at 37°C in 96-well microculture plates. 5.25.EGFP.Luc.M7 cells (kindly provided by Dr. Nathaniel Landau (Brandt *et al.*, *J. Biol. Chem.* 277:17291-9 (2002)) suspended in RPMI with 12% fetal bovine serum and 25 μg/ml DEAE-dextran at $7.5x10^4$ cells per well were then added and incubated for 3–4 days. These cells express luciferase and GFP upon HIV infection. Luciferase was measured using a commercial kit (Bright-Glo, Promega, Madison, WI) and a Fluoroskan luminometer. The titer was calculated as the dilution at which relative light units (RLU) were 50% that of virus-only wells. For the large panel of viruses, sera were assayed at a 1:10 dilution and the percent reduction in relative light units (RLU) was calculated in relation to a pre-immunization serum sample for each rabbit. A reduction of 50% is considered significant in this assay.

Results

[0304] *Method validation:* First, the ability of likelihood-based phylogenetic methods to accurately predict ancestral sequences was evaluated using both a simulation and an experimental approach. HIV-1 is widely considered to have a star-like phylogeny with all external branches radiating from the same central point on the tree. However, there is at least some level of substructure in the phylogeny (*i.e.*, some short internal branches at the base of the tree). In terms of tree space, the realized HIV-1 tree (see Figure 45B) is somewhere between a true star and what can be called a caterpillar tree (*i.e.*, long external branches with short internal branches) (Figure 45a). To estimate the accuracy of ancestral state reconstruction, 100 simulations were performed with a known ancestor at each of the two extremes of the star-caterpillar continuum. For the caterpillar tree, the MRCA was estimated with 95.4% accuracy. For the star phylogeny, the MRCA was estimated with 98.2% accuracy. Since the true HIV phylogeny falls somewhere between the caterpillar and star structures, the MRCA reconstruction accuracy was estimated to be in the range of 95-98%.

[0305] Viral gene sequences taken from 4 rhesus macaques infected with a molecularly defined strain of SIV, SIVmac251-BK28 (Kornfeld *et al.*, *Nature* 326:610-3 (1987) were used in an effort to assess the reconstruction of the infecting viral sequence. From a phylogenetic tree of 1.1kb *env* gene fragments taken between 1–3 years post infection, 99.8% of the infecting viral sequence (373 of 376 amino acids), including 98.2% of the 170 variable sites, were accurately predicted.

[0306] *Derivation of subtype B ancestor:* Using the same procedures, an ancestral nucleotide sequence of the envelope gene of HIV-1 subtype B, the most extensively evaluated HIV-1 subtype to date, was derived using the phylogeny shown in Figure 45b. This ancestral sequence (AN1-EnvB) produced an open reading frame encoding a complete, 884-amino acid gp160 gene product. The amino acid distances between the ancestral sequence and the natural subtype B strains used to estimate it were 12.3% on average (range: 8.0-21.0%) while these sequences were 17.3% different from each other (range: 13.3-23.2%). At 884 amino acids, the encoded protein is long relative to most sequences; much of the additional length is in the variable loops. Also unusual is the number of putative sites for N-linked glycosylation, also called sequins: the average is 25, while AN1-EnvB has 35. Based on features V3 sequence features, AN1-EnvB is predicted to use CCR5 as a coreceptor

[0307] *In vitro validation of ancestor protein expression.* The functional characteristics of the ancestral *env* sequence were evaluated. The deduced sequence was re-engineered to reflect the dominant patterns of human codon usage, without changing any of the encoded amino acids, and then chemically synthesized the gene and cloned it into a mammalian expression vector pVR1012. Transfection of this plasmid into COS-7 or 293T cells resulted in the production of high levels of viral protein gp160 and its cleavage products, gp120 and gp41 – comparable to levels detected following transfection of the widely employed, humanized Env-gp160 gene of SHIV89.6P (Barouch *et al.*, *Science* 290:486-92 (2000)). Control of viremia and prevention of clinical AIDS in rhesus monkey and higher than that of non-optimized HIV-1 89.6 Env. A truncated gp140 form, which is expected to retain the oligomeric properties of gp160 but is secreted, was also generated and expressed and shown to bind soluble CD4 in an ELISA assay. Western blot analysis of 293T cells transfected with either the gp160 or gp140 form showed that the expressed proteins were of the predicted sizes and were recognized by anti-HIV antibodies. Cell lysates contained both gp140 and gp160, while only the gp140 form appeared in the medium, showing that it was secreted as expected. Lysates of cells transfected with gp160 showed bands for gp41 and gp120, indicating that Env was cleaved to its mature form intracellularly. The cleavage was not complete, as full-length gp160 was also visible. The apparent molecular weight and somewhat diffuse nature of the bands suggested that the AN1-EnvB proteins were highly glycosylated.

[0308] Expression of gp160 in transfected cells was confirmed by an immunofluorescence assay. Cell-cell fusion occurred when GHOST-CCR5 cells (Cecilia *et al.*, *J. Virol.* 72:6988-96 (1998)) expressing the HIV-1 receptors CD4 and CCR5 were transfected with AN1-EnvB gp160. In contrast, no fusion occurred when GHOST-CXCR4 cells expressing CD4 plus CXCR4 were transfected, although Env was expressed in these cells. These results indicated that AN1-EnvB used only CCR5, as predicted by its sequence, and that it was folded properly and fully functional for directing membrane fusion. To confirm the functionality of AN1-EnvB, pVR1012AN1env (encoding full-length gp160) was used to complement the *env*-defective genome Q23Δenv(Long *et al.*, *AIDS Res. Hum. Retroviruses* 18:567-76 (2002)). The resulting pseudotyped virus particles were able to infect permissive cMAGI cells. Concentrated virions had a titer of 1.1×10^4 infectious particles/ml on cMAGI cells.

[0309] *Immunogenicity of AN1-EnvB:* To assess the humoral immune response potentially elicited by An1-EnvB, rabbits were immunized with DNA encoding gp160 and gp140 versions of the protein via Gene Gun followed by boosting with recombinant AN1-EnvB gp120 protein made in CHO cells. As controls, additional groups were immunized with a codon-optimized Env gp140 from a primary HIV-1 isolate, HIV-SF162 and recombinant SF162 gp120, or with the empty vector and adjuvant. All Envelope plasmids elicited Envelope-specific antibody binding titers in rabbits up to 1:1,000,000. Titers increased with protein boosting.

[0310] *Virus neutralization:* Antibody neutralization was measured using the cMAGI assay (Chackerian *et al.*, *J. Virol.* 71:3932-9 (1997); Kimpton *et al.*, *J. Virol.* 66:2232-9 (1992)) with five heterologous subtype B, primary R5 HIV-1 isolates SF162, 92US657, 92TH014, 91US056, and 92HT593, and one subtype C primary isolate, 93IN101 (Table 13). As a positive control, HIV^+ sera pooled from several local subtype B-infected patients was included. All isolates tested were neutralized by this pool, but with differing sensitivities. After four DNA immunizations, all gp140-SF162-immunized rabbit sera had neutralizing activity against the homologous virus. Six of eight AN1-EnvB immunized rabbit sera also neutralized HIV-1 SF162, which in this case was heterologous. Seven of these eight rabbits had 50% neutralization titers of at least 1:8 against 92US657; all eight achieved 90% neutralization against 92TH014 after five immunizations (geometric mean titer=16). In contrast, sera against SF162 protein were less broadly reactive, with neutralization titers of at least 1:8 detected against two of five heterologous viruses (in one of four animals against 93IN101 and in two of four animals against 92US657). Furthermore, HIV 92TH014 and 92US657 were more effectively neutralized by the AN-1-EnvB-immunized rabbit sera compared with Env SF162 sera, with high titers at the 90% cutoff against 92TH014. Two of the six HIV-1 isolates tested were only very weakly or not neutralized by sera from any of the Env-immunized rabbits (92HT593 and 91US056), although the human HIV-positive serum control neutralized both. Subtype C 93IN101 was weakly neutralized by sera from at most one animal in each group.

[0311] These data were extended in a second neutralization assay using 5.25.EGFP.Luc.M7 cells. Strong neutralizing activity against HIV-1 SF162 was found in all Env-immunized rabbits, confirming the results in the cMAGI assay. This activity increased over the course of

vaccinations. Titers increased with each of the first two immunizations and then declined, as has been seen with vaccinations in primates, while subsequent boosts resulted in sustained titers at a threshold level that was not additionally boosted. Levels of neutralizing antibody against HIV-SF162 were similar in each of the groups receiving homologous Env-SF162, or heterologous AN1-Env-B gp140 or AN1-Env-B-gp160. Inhibition of primary isolate Bx08 was found in four of eight AN1-EnvB- and in one of four SF162-immunized rabbits. However, none of the AN1-Env B-immunized rabbit sera neutralized BaL and JR-FL, whereas sera raised against Env-SF162 neutralized BaL and in one case, JR-FL. In summary, at least five of the eight primary (all heterologous) clade B viruses tested were neutralized by AN1-Env B sera (Table 14).

Discussion

[0312] Artificial HIV gene sequences, such as ancestors and consensus can be more effective vaccines and reagents than natural isolates. Such sequences can be engineered to contain more of the conserved features and epitopes found in primary isolates than any one isolate would have. This concept has been put into practice by predicting, synthesizing, and evaluating an ancestor of the HIV-1 clade B *env* gene. The AN1-EnvB protein can be stably expressed in mammalian cell lines, was glycosylated, bound CD4, directed cell-cell fusion, used CCR5 but not CXCR4, and complemented *env*-defective genomes. This functional protein elicited high titers of Env-binding antibodies in vaccinated mice and rabbits. After DNA priming and protein boosting, immunized rabbit sera neutralized a variety of primary isolates at a low level. gp140 constructs gave slightly higher titers (binding and neutralizing) than gp160, but this difference was not statistically significant.

[0313] The neutralization profile of AN1-EnvB-vaccinated rabbit sera supports the use of ancestral viral sequences as immunogenic compositions. In an effort to examine breadth of neutralization in an impartial manner, coded sera were tested using two different assays, performed in two different laboratories, using panels of laboratory and primary viruses with a single isolate in common, HIV-SF162. (No standard panel of HIV-1 primary isolates has been established for comparisons between laboratories to date.) Results in both laboratories indicated that sera from the AN1-EnvB-immunized rabbits could neutralize heterologous HIV-1 strains to a significant degree. Three of five subtype B viruses were neutralized by a majority of sera using the cMAGI assay and three of five subtype B viruses were neutralized

by at least some of the sera in the M7-luc assay. Several rabbits showed low-level responses against one virus from subtype C, but none neutralized another C or the E subtype viruses tested. This profile was similar to that of sera from rabbits immunized with Env from HIV-1 SF162. Indeed, the neutralization of HIV-1 SF162 itself was similar for sera from rabbits given either immunogen, even though this isolate is homologous to the SF162 immunogen but heterologous to the AN1-EnvB immunogen. These data imply that the artificial sequence was at least as immunogenic as this natural isolate.

[0314] Several recent vaccine experiments have shown that DNA expressing unmodified HIV-1 Envelopes can elicit low-level antibodies that neutralize the homologous primary isolate at the 50% level. Low titer cross-reactive neutralizing antibodies were elicited using modified Envelope DNA vaccines or recombinant proteins. The codon-optimized vectors used here elicited high titers of binding antibodies with DNA vaccination alone. Responses to DNA vaccines can be effectively boosted with recombinant viral vectors or recombinant glycoproteins. The results reported here support the use of DNA vaccines alone in demonstrating low-level cross-clade neutralizing antibodies. Immunization of the DNA-primed rabbits with homologous gp120 subunit protein in adjuvant resulted in boosting of binding antibody titers against HIV-SF162. However, gp120 failed to boost heterologous (*i.e.*, from a different HIV subtype) neutralizing antibody responses significantly. The gp140 and gp160 DNA vaccines presented native envelope conformations *in vivo* that were boosted poorly or not at all by adjuvated gp120 monomer proteins, raising responses only to the most neutralization-sensitive primary virus, HIV-SF162. Although native gp120 can bind conformation-dependent broad neutralizing antibodies, elicit neutralizing antibodies and boost DNA priming, there is evidence that use of oligomeric envelope proteins may be more effective in boosting broader responses.

Table 11. Sequences used for phylogenetic reconstruction

Sequence name	Accession number	Subtype
HIVELICG	K03454	D
HIVNDK	M27323	D
HIVZ2Z6	M22639	D
87USSG3X	L02317	B
88USWR27	U26546	B
89SP061	AJ006287	B
AUC18	AF042102	B

AUC18MBC	U37270	B
AUMBC200	AF042100	B
AUMBC925	AF042101	B
AUMBCC18B	AF042106	B
AUMBCC54	AF042103	B
AUMBCC98	AF042104	B
AUMBCD36	AF042105	B
CNRL42CG	U71182	B
D31	U43096	B
HIVBH102	M15654	B
HIVCAM1	D10112	B
HIVF12CG	Z11530	B
HIVHAN2	U43141	B
HIVJC16	AF049494	B
HIVJRCSF	M38429	B
HIVJRFL	U63632	B
HIVLAICG	K02013	B
HIVMCK1	D86068	B
HIVMN	M17449	B
HIVNL43	M19921	B
HIVNY5CG	M38431	B
HIVOYI	M26727	B
HIVPV22	K02083	B
HIVRF	M17451	B
HIVSF2CG	K02007	B
HIVWEAU160	U21135	B
HIVYU10X	M93259	B
HIVYU2X	M93258	B
MANC	U23487	B
NLACH320A	U34603	B
NLACH320B	U34604	B
US89.6	U39362	B
USAD8	U19647	B
USDH123	AF069140	B

Table 12. Parameters for evolutionary models used for phylogenetic reconstructions.

	HIV-1-B *ENV*	SIV *env*
Evolutionary Model Class[a]	GTR+I+G	GTR+I+G
RC ↔ A[b]	1.971	1.8418
RG ↔ A	4.315	10.9127
RT ↔ A	0.8539	0.5803
RC ↔ G	1.021	0.9283
RC ↔ T	3.826	6.8499
*f*A[c]	0.34857	0.3516
*f*C	0.17125	0.1642
*f*G	0.23694	0.2296
*f*T	0.24324	0.2546
Proportion of assumed invariant sites	0.26035	0.471
α[d]	0.50434	0.9477

[a] GTR+I+G: General time-reversible model with a proportion of invariant sites and gamma-distributed site-to-site rate variation.

[b] Rate parameters of the symmetric substitution rate matrix: RX ↔ Y is the rate of substitution of nucleotide X by nucleotide Y (or Y by X) scaled to RG ↔ T = 1.

[c] *f*X: Equilibrium frequency of nucleotide X.

[d] α: Shape parameter of the gamma distribution.

Table 13 Neutralization of HIV-1 primary isolates in the cMAGI assay

HIV isolate		SF162				92US657		92TH014		92HT593	91US056		93IN101
Immunization		4	4	5	5	5	5	5	5	5	5	5	5
Group	No.	50%	75%	50%	75%	50%	75%	50%	90%	50%	50%	75%	50%
	32N	**22**	**15**	**16**	<8	<8	<8	<8	<8	<			
SF162										4	<4	<4	<8
Gp140	91N	**22**	**18**	<8	<8	<8	<8	***55**	<8	<4	<4	<4	<8
	9335L	**9**	**9**	<8	<8	**9**	<8	<8	<8	<4	<4	<4	<8
	9578L	**40**	**13**	**18**	**9**	**14**	<8	nt	nt	<4	<4	<4	**8**
Ancestral	168N	<4	<4	<8	<8	**20**	**12**	**40**	**16**	<4	**16**	<4	<8
Gp140	245N	**11**	<4	<8	<8	**32**	**16**	**88**	**48**	<4	<4	<4	<8
	247N	**5**	<4	<8	<8	**10**	<8	**32**	**20**	<4	<4	<4	<8
	9331L	**6**	<4	**8**	<8	**28**	**13**	***52**	***32**	<4	<4	<4	**8**
Ancestral	93N	**6**	**4**	<8	<8	**18**	**8**	**40**	**16**	<4	**4**	<4	<8
Gp160	175M	**6**	**4**	**8**	<8	**12**	<8	**24**	**10**	<4	**4**	<4	<8
	9599L	<4	<4	<8	<8	**9**	<8	**44**	**24**	<4	<4	<4	<8
	9896L	**8**	**4**	<8	<8	<8	<8	**40**	**16**	<4	<4	<4	nt
Control	9827L	<4	<4	<8	<8	**8**	<8	**15**	<8	<4	<4	<4	<8
pcDNA	8848L	<4	<4	<8	<8	<8	<8	**8**	<8	<4	<4	<4	<8
HIV+ serum pool		nt	nt	**2500**	**500**	**1100**	**300**	**300**	**100**	**200**	**500**	**300**	**400**

Individual pre-immunization sera were tested for isolates SF162, 92US657, 92TH014, and 93IN101; pooled sera were tested for 92HT593 and 91US056. All were below 50% at the indicated dilution, except as noted; for samples marked * the pre-bleed titer was subtracted. nt, not tested.

Table 14

Neutralization of HIV-1 by sera from immunized rabbits – Week 80 (post-8th vaccination)

		% Reduction in RLU of:						
Rabbit	Immunogen	**Bal clade B**	**JR-FL clade B**	**Bx08 clade B**	**6101 clade B**	**0692 clade B**	**S080 clade C**	**CM244 clade E**
91N	SF162	**62**	32	**68**	43	**53**	24	0
9335L	SF162	**66**	38	**66**	25	32	0	13
32N	SF162	0	0	**65**	13	35	0	0
175N	AN-gp160	**71**	36	**61**	1	**62**	0	27
93N	AN-gp160	36	0	45	0	35	0	0
9599L	AN-gp160	35	0	**59**	0	44	0	0
9896L	AN-gp160	0	0	0	5	0	0	0
245N	AN-gp160	30	3	45	8	40	0	2
247N	AN-gp160	0	0	25	0	41	0	33
9331L	AN-gp160	0	0	14	0	1	0	0
168N	AN-gp160	**69**	34	**73**	0	**65**	0	0
8848L	Control	0	16	29	26	0	0	0
9827L	Control	**51**	0	19	10	0	0	0

[1]*Samples were assayed at a 1:10 dilution in triplicate. Values are the % reduction in relative luciferase units compared to the corresponding week 5 sample (05-16-02). Values ≥50% neutralization are shown in boldface type.*

[0315] From the foregoing, it will be appreciated that, although specific embodiments of the invention have been described herein for the purpose of illustration, various modifications may be made without deviating from the spirit and scope of the invention. All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to one of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.